MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS

②

MARYLAND

COLLEGE PARK CAMPUS

EIGENVALUE PROBLEMS

I. Babuška
Institute for Physical Science and Technology
and
Department of Mathematics
University of Maryland
College Park, Maryland 20742

and

J. Osborn
Department of Mathematics
University of Maryland
College Park, Maryland 20742

BN-1066

DTIC
ELECTE
S
NOV 2 0 1987
D
H

June 1987

INSTITUTE FOR PHYSICAL SCIENCE AND TECHNOLOGY

33

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
BN-1066	A188 069	

4. TITLE *(and Subtitle)*	5. TYPE OF REPORT & PERIOD COVERED
Eigenvalue Problems	Final life of the contract
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
I. Babuska[1] and J. Osborn	[1]ONR N00014-85-K-0169

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Institute for Physical Science and Technology University of Maryland College Park, MD 20742	

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
Department of the Navy Office of Naval Research Arlington, VA 22217	June 1987
	13. NUMBER OF PAGES 235

14. MONITORING AGENCY NAME & ADDRESS(*If different from Controlling Office*)	15. SECURITY CLASS. *(of this report)*
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*
The report is a chapter in the "Handbook of Numerical Analysis", edited by P.G. Ciarlet and J.L. Lions and gives a comprehensive survey of the state of the art in the numerical eigenvalue analysis.

EIGENVALUE PROBLEMS

I. Babuška[*] and J. Osborn[**]

To appear in <u>Handbook of Numerical Analysis</u>,
edited by P.G. Ciarlet and J.L. Lions, and
to be published by North-Holland, Amsterdam.

[*] Institute for Physical Science and Technology and Department of
Mathematics, University of Maryland, College Park, MD 20742. The
work of this author was partially supported by the Office of Naval
Research under Contract N00014-85-K-0169 and by the National
Science Foundation under Grant DMS-85-16191.

[*] Department of Mathematics, University of Maryland, College Park,
MD 20742. The work of this author was partially supported by the
National Science Foundation under Grant DMS-84-10324.

CHAPTER I. INTRODUCTION AND PRELIMINARIES

Section 1. Examples of Eigenvalue Problems

In this section we present several model eigenvalue problems
arising in physics and engineering. Specifically, we will discuss
briefly some important physical interpretations of eigenvalues and
eigenfunctions. Some of the model problems we discuss here will
serve as illustrative examples in connection with the approximation
methods considered in Chapter III. We will attempt to provide a
clear understanding of the fundamental ideas, but will not present
a detailed treatment. For a more complete discussion of the mate-
rial in this section we refer to Courant-Hilbert [1953].

A. One Dimensional Problems

The Longitudinal Vibration of an Elastic Bar

We are interested in studying the small, longitudinal vibra-
tions of a longitudinally loaded, elastically supported, elastic
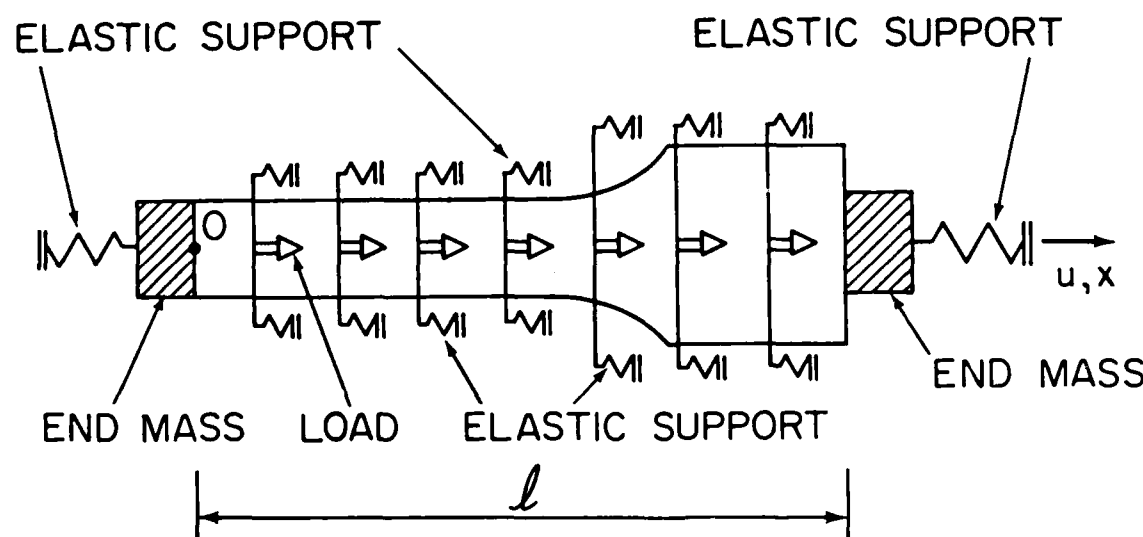bar with masses attached to its ends. The bar is shown in Figure
1.1.

Figure 1.1. Elastic Bar.

We now derive the governing differential equation and boundary conditions for the problem. First we consider the static problem. Suppose

$f(x)$, $0 < x < \ell$, represents the external longitudinal load, with positive $f(x)$ denoting a force directed to the right,

$u(x)$, $0 < x < \ell$, denotes the displacement of the cross-section of the bar originally at x, with positive $u(x)$ denoting the displacement to the right, so that the position of a point originally at x is $x + u(x)$,

$\varepsilon(x)$, $0 < x < \ell$, denotes the strain in the x-direction, i.e., the relative change in the length of the fibers in the bar ($\varepsilon(x)$ will be positive if it describes extension),

$\sigma(x)$, $0 < x < \ell$, denotes the normal stress in the cross-section at x, i.e., the force per unit area exerted by the portion of the bar to the right of x on the portion to the left of ($\sigma(x)$ will be positive if it describes tension),

$A(x)$, $0 < x < \ell$, denotes the area of the cross-section at x,

$E(x)$, $0 < x < \ell$, denotes the modulus of elasticity of the bar at x,

$F(x)$, $0 < x < \ell$, denotes the internal force acting on the cross-section at x, i.e., the force exerted by the portion of the bar to the right of x on the portion to the left, with positive $F(x)$ denoting a force directed to the right,

$\rho(x)$, $0 < x < \ell$, denotes the load due to the (continuous) elastic support, which is assumed to be of the form

$$\rho(x) = -c(x)u(x),$$

where $c(x) > 0$ is the spring constant of the support (the

3

negative sign indicates that the force is directed opposite to the displacement), and

$m(x)$, $0 < x < \ell$, denotes the specific mass at x, i.e., the mass per unit volume at x.

The strain $\varepsilon(x)$ and the displacement $u(x)$ are related by

$$\varepsilon(x) = \frac{du}{dx}(x).$$

This relation is valid for small displacements, i.e., when $|\varepsilon(x)| \ll 1$. The relation between stress and strain is described by the constituitive law of the material. We are assuming the linear relation given by Hooke's Law:

$$\sigma(x) = E(x)\varepsilon(x).$$

Thus, since $F(x) = \sigma(x)A(x)$, we have

$$F(x) = A(x)E(x)\varepsilon(x)$$

$$= A(x)E(x)\frac{du}{dx}(x).$$

Now the equilibrium condition for the bar is

$$\frac{dF}{dx}(x) + f(x) + \rho(x) = 0,$$

which, with the use of the relations discussed above, can also be written as

(1.1) $-\dfrac{d}{dx}\left(A(x)E(x)\dfrac{du}{dx}(x)\right) + c(x)u(x) = f(x),\ 0 < x < \ell.$

This is the governing differential equation.

We consider the three most important types of boundary conditions.

Dirichlet Type

(1.2a) $u(0) = a_1,\ u(\ell) = a_2$

4

Here the displacements of the end points of the bar are given.

Neumann Type

$$(1.2b) \qquad -F(0) = -(AE\tfrac{du}{dx})(0) = b_1, \quad F(\ell) = (AE\tfrac{du}{dx})(\ell) = b_2$$

Here the forces at the ends of the bar are given. The different signs at 0 and ℓ are used to express the outer normal derivative at the ends of the bar.

Newton Type

$$(1.2c) \quad -(AE\tfrac{du}{dx})(0) + \gamma_1 u(0) = c_1, \quad (AE\tfrac{du}{dx})(\ell) + \gamma_2 u(\ell) = c_2,$$

$$\text{where} \quad \gamma_1, \gamma_2 > 0$$

Here γ_2 is the spring constant of a spring attached to the bar at $x = \ell$ and $-\gamma_2 u(\ell)$ is the force exerted on the right end of the bar by the spring. We are thus specifying the sum of the internal force and the spring force on the right end of the bar. The condition at $x = 0$ has a similar interpretation.

(1.1) together with one of (1.2a,b,c) determine the displacement $u(x)$ in the static case. We now turn to the <u>dynamic case</u>.

We assume the external load depends on the time t and is represented by $f(x,t)$ and suppose a_i, b_i, c_i in the boundary conditions depend on t : $a_i = a_i(t)$, $b_i = b_i(t)$, $c_i = c_i(t)$, $i = 1,2$. We further suppose the bar is subject to a damping force represented by R. If $u = u(x,t)$ is the displacement at time t, then from Newton's 2nd law we have

(1.3)

$$-\frac{\partial}{\partial x}(A(x)E(x)\frac{\partial u}{\partial x}(x,t)) + c(x)u(x,t) = f(x,t) - m(x)A(x)\frac{\partial^2 u}{\partial t^2}(x,t) - R,$$

$$0 < x < \ell, \ t > 0.$$

5

We next give the boundary conditions in the dynamic case. The Dirichlet conditions are nearly the same as in the static case, while the Neumann and Newton conditions require modification because of the forces exerted on the ends of the bar by the attached masses.

Dirichlet Type

(1.4a) $u(0,t) = a_1(t), \; u(\ell,t) = a_2(t), \; t \geq 0$

Neumann Type

(1.4b)
$$
\begin{cases}
(-AE\dfrac{\partial u}{\partial x})(0,t) = -m_1\dfrac{\partial^2 u}{\partial t^2}(0,t) + b_1(t) \\[4mm]
(AE\dfrac{\partial u}{\partial x})(\ell,t) = -m_2\dfrac{\partial^2 u}{\partial t^2}(\ell,t) + b_2(t), \; t \geq 0,
\end{cases}
$$

where m_1 and m_2 are the masses attached to the left and right ends of the bar, respectively

Newton Type

(1.4c)
$$
\begin{cases}
(-AE\dfrac{\partial u}{\partial x})(0,t) + \gamma_1 u(0,t) = -\, m_1\dfrac{\partial^2 u}{\partial t^2}(0,t) + c_1(t) \\[4mm]
(AE\dfrac{\partial u}{\partial x})(\ell,t) + \gamma_2 u(\ell,t) = -\, m_2\dfrac{\partial^2 u}{\partial t^2}(\ell,t) + c_2(t), \; t \geq 0
\end{cases}
$$

We remark that we can impose boundary conditions of different types at the two ends. For example, we could impose a Newton type condition at 0 and a Dirichlet type at ℓ.

Finally in this (dynamic) case we need to impose initial conditions. We specify the initial position and velocity:

(1.5)
$$
\begin{cases}
u(x,0) = r_1(x) \\[2mm]
\dfrac{\partial u}{\partial t}(x,0) = r_2(x), \; 0 < x < \ell
\end{cases}
$$

Consider now equations (1.3), with $f = R = 0$, and one of the conditions (1.4a,b,c), with $a_1 = a_2 = b_1 = b_2 = c_1 = c_2 = 0$. If we seek <u>separated</u> <u>solutions</u> of the form

$$u(x,t) = v(x)w(t),$$

in which the spatial variable x and the temporal variable t are separated, from (1.3) we find that

$$[-\frac{d}{dx}(A(x)m(x)\frac{dv}{dx}(x)) + c(x)v(x)]w(t) = -m(x)A(x)v(x)\frac{d^2w}{dt^2}(t)$$

or

$$(1.6) \quad \frac{-\frac{d}{dx}(A(x)E(x)\frac{dv}{dx}(x)) + c(x)v(x)}{m(x)A(x)v(x)} = \frac{-\frac{d^2w}{dt^2}(t)}{w(t)}, \quad 0 < x < \ell, \ t \geq 0.$$

Imposing the boundary conditions (1.4a,b,c) on $u = vw$ we find

$$(1.7a) \qquad v(0)w(t) = 0, \ v(\ell)w(t) = 0, \ t \geq 0$$

$$(1.7b) \quad \begin{cases} \dfrac{-(AE\frac{dv}{dx})(0)}{m_1 v(0)} = \dfrac{-\frac{d^2w}{dt^2}(t)}{w(t)} \\[20pt] \dfrac{(AE\frac{dv}{dx})(\ell)}{m_2 v(\ell)} = \dfrac{-\frac{d^2 w}{dt^2}(t)}{w(t)} \ , \ t \geq 0, \end{cases}$$

$$(1.7c) \quad \begin{cases} \dfrac{-(AE\frac{dv}{dx})(0) + \gamma_1 v(0)}{m_1 v(0)} = \dfrac{-\frac{d^2w}{dt^2}(t)}{w(t)} \\[20pt] \dfrac{(AE\frac{dv}{dx})(\ell) + \gamma_2 v(\ell)}{m_2 v(\ell)} = \dfrac{-\frac{d^2w}{dt^2}(t)}{w(t)}, \quad t > 0. \end{cases}$$

It is immediate that both sides of equation (1.6) equal a

7

constant, which we denote by λ. We are thus led to seek a number

λ and a function $v(x) \neq 0$ so that

(1.8) $-\dfrac{d}{dx}(A(x)E(x)\dfrac{dv}{dx}(x)) + c(x)v(x) = \lambda m(x)A(x)v(x), \; 0 < x < \ell.$

From (1.7a,b,c) we get boundary conditions for v:

(1.9a) $v(0) = v(\ell) = 0,$ (Dirichlet type)

(1.9b) $\begin{cases} -(AE\dfrac{dv}{dx})(0) = \lambda m_1 v(0) \\[2ex] (AE\dfrac{dv}{dx})(\ell) = \lambda m_2 v(\ell), \end{cases}$ (Neumann type)

(1.9c) $\begin{cases} -(AE\dfrac{dv}{dx})(0) + \gamma_1 v(0) = \lambda m_1 v(0) \\[2ex] (AE\dfrac{dv}{dx})(\ell) + \gamma_2 v(\ell) = \lambda m_2 v(\ell), \end{cases}$ (Newton type).

The problem of finding λ and $v(x) \neq 0$ satisfying (1.8) and a

boundary condition (1.9) of Dirichlet, Neumann, or Newton type is

called an _eigenvalue_ problem. λ is called an _eigenvalue_ and

$v(x)$ a corresponding _eigenfunction_, or _eigenvector_, of the prob-

lem, and (λ,v) is often called an _eigenpair_. If λ is present

in one or both of the boundary conditions, the problem is referred

to as a _Steklov-type_ eigenvalue problem.

For the sake of definiteness, let us suppose we have a Newton

type boundary condition at 0 and a Dirichlet type at ℓ, and

further assume that $m_1 = 0$. Thus we are considering the initial-

boundary value problem

(1.3') $-\dfrac{\partial}{\partial x}(AE\dfrac{\partial u}{\partial x}) + cu = -mA\dfrac{\partial^2 u}{\partial t^2}, \; 0 < x < \ell, \; t > 0$

(1.4c') $-(AE\dfrac{\partial u}{\partial x})(0,t) + \gamma_1 u(0,t) = 0,$

(1.4a') $u(\ell,t) = 0, \; t > 0$

8

$$(1.5') \qquad \begin{cases} u(x,0) = \gamma_1(x) \\ \dfrac{\partial u}{\partial t}(x,0) = \gamma_2(x), \ 0 < x < \ell. \end{cases}$$

The corresponding eigenvalue problem is

$$(1.10) \qquad \begin{cases} -\dfrac{d}{dx}\left(AE\dfrac{du}{dx}\right) + cv = \lambda m A v, \ 0 < x < \ell \\ -\left(AE\dfrac{du}{dx}\right)(0) + \gamma_1 v(0) = 0 \\ v(\ell) = 0. \end{cases}$$

It is known that problems of this type have a sequence of eigenvalues

$$(1.11) \qquad 0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow +\infty$$

and corresponding eigenfunction

$$(1.12) \qquad v_1(x), \ v_2(x), \ \ldots \ .$$

The eigenfunctions satisfy

$$(1.13) \qquad \int_0^\ell m(x)A(x)v_i(x)\bar{v}_j(x)dx = \delta_{ij},$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \ne j$, i.e., they are orthonormal; in addition they are complete in L_2, i.e., any function $h(x) \in L_2$ can be written as

$$(1.14) \qquad h(x) = \sum_{j=1}^\infty c_j v_j(x),$$

where

$$(1.15) \qquad c_j = \int_0^\ell m A h \bar{v}_j dx$$

and the convergence is in the L_2-norm. Regarding (1.11) - (1.15),

see (4.10) - (4.14).

Corresponding to each λ_j we solve

$$(1.16) \qquad \frac{d^2w}{dt^2}(t) + \lambda_j w(t) = 0, \ t > 0$$

(cf. (1.6)), obtaining

$$w(t) = w_j(t) = a_j \sin \sqrt{\lambda_j}(t+\theta_j),$$

where a_j and θ_j are arbitrary. Thus the separated solutions are given by

$$(1.17) \qquad a_j v_j(x) \sin \sqrt{\lambda_j}(t+\theta_j), \ j = 1,2,\ldots$$

It is immediate that

$$(1.18) \qquad u(x,t) = \sum_{j=1}^{\infty} a_j v_j(x) \sin \sqrt{\lambda_j}(t+\theta_j)$$

is a solution of (1.3'), (1.4c'), (1.4a') for abitrary a_j and θ_j, provided the series converges appropriately. It remains to satisfy the initial conditions (1.5'). For this, a_j and θ_j must satisfy

$$u(x,0) = \sum_j a_j \sin \sqrt{\lambda_j}\theta_j v_j(x) = \tau_1(x),$$

$$\frac{\partial u}{\partial t}(x,0) = \sum_j a_j\sqrt{\lambda_j} \cos \sqrt{\lambda_j}\theta_j v_j(x) = \tau_2(x).$$

From the complete orthonormality of the $v_j(x)$ we see that these two equations uniquely determine a_j and θ_j. Thus (1.18), with this choice for a_j and θ_j, is the unique solution of (1.3'), (1.4c'), (1.4a'), (1.5').

The simple motions given in (1.17) are called the eigenvibrations of (1.3'), (1.4c'), (1.4a'). All the points x of the

j^{th} eigenvibrations vibrate with the same (circular) frequency (defined to be the number of vibrations per 2π seconds) and phase displacement $\sqrt{\lambda_j}\theta_j$ and the point x vibrates with amplitude proportional to $v_j(x)$. Thus $\sqrt{\lambda_j}$ is the frequency with which the jth eigenvibration vibrates and $v_j(x)$ gives the basic shape of the eigenvibration. The amplitude factor a_j and θ_j are determined by the initial position and velocity of the eigenvibration, whereas λ_j and $v_j(x)$ are determined by the physical process itself, as represented by (1.3'), (1.4c'), and (1.4a'). We have seen that any motion of (1.3'), (1.4c'), (1.4a') can be written as a sum or superposition of eigenvibrations.

So far we have been dealing with free vibrations, i.e., we have assumed $f(x,t)$ and R in (1.3) are zero. Now we briefly consider the case when $f \neq 0$ and $R = 0$, i.e., the case of forced vibrations. If we write

$$f(x,t) = \sum_{j=1}^{\infty} f_j(t)\bar{v}_j(x)m(x)A(x),$$

then we easily see that $u(x,t) = \sum_{j=1}^{\infty} {}_j a_j(t)v_j(x)$ is a solution if

$$a_j''(t) + \lambda_j a_j(t) = f_j(t).$$

If, now, $f_j(t) = \sin\sqrt{\lambda_j}(t+\theta_j)$, then we see that $a_j(t)$, and hence $u(x,t)$, will be unbounded as $t \to \infty$. This phenomena is called <u>resonance</u> and f is called a <u>resonant load</u>; the <u>resonant frequencies</u> are $\sqrt{\lambda_j}$, $j = 1,2,\ldots$.

The damping term R could be defined in various ways. For example, we could take R to be $\mu\frac{\partial u}{\partial t}$, for a constant μ, which

11

would lead to a term of the form $\mu\dfrac{\partial u}{\partial t}$ in equation (1.3).

Eigenvalue problems similar to (1.8) and (1.9) or (1.10) arise in a number of other situations. We now briefly mention some of them.

The Transverse Vibration of a String

We are interested here in the small, transverse vibration of a homogeneous string that is stretched between two points a distance ℓ apart. Gravity is assumed to be negligible and the particles of the string are assumed to move in a plane. We denote the density of the string by r and the tension by p. We restrict our attention to the case of free vibrations.

If the particles of the string are identified with the numbers $0 \le x \le \ell$ and if $u(x,t)$ denotes the vertical displacement of the particle x at time t, then u satisfies

$$(1.19) \quad \begin{cases} -p\dfrac{\partial^2 u(x,t)}{\partial x^2} = -r\dfrac{\partial^2 u(x,t)}{\partial t^2}, \ 0 < x < \ell, \ t > 0 \\ u(0,t) = u(\ell,t), \ t \ge 0. \end{cases}$$

We see that (1.19) is a very special case of (1.3) and (1.4a). The associated eigenvalue problem is

$$(1.20) \quad \begin{cases} -c^2 v''(x) = \lambda v(x), \ 0 < x < \ell \\ v(0) = v(\ell) = 0, \end{cases}$$

where $c^2 = p/r$. It is easily seen that the eigenvalues and eigenfunctions of (1.20) can be given explicitly; they are

$$(1.21) \quad \lambda_k = \frac{k^2 c^2 \pi^2}{\ell^2}$$

and

12

$$(1.22) \qquad v_k(x) = \sqrt{2/\ell} \, \sin \frac{k\pi x}{\ell}, \quad k = 1, 2, \ldots \,.$$

The entire discussion of the elastic bar — i.e., the discussion of separation of variables, of eigenvalues and eigenfunctions, and of eigenvibrations — applies to this problem. We note that it is possible to find the eigenvalues and eigenfunctions explicitly only in very special situations, roughly, just in the case of eigenvalue problems for differential equations with constant coefficients in one dimension. In general, one must resort to approximation methods. The discussion of such methods is the main topic of this article.

Characterization of the Optimal Constant in the Poincaré Inequality

The Poincaré inequality states that there is a constant C such that

$$(1.23) \qquad \int_0^\ell [u(x)]^2 dx \leq C \int_0^\ell [u'(x)]^2 dx$$

for all functions $u(x)$ having a square integrable first derivative and vanishing at 0 and ℓ. Let us consider the problem of finding the minimal constant C. We are thus interested in

$$(1.24) \qquad C = \sup_{\substack{u \\ u(0)=u(\ell)=0}} \frac{\int_0^\ell u^2 dx}{\int_0^\ell (u')^2 dx} \,.$$

Using the elementary methods of the calculus of variations we find that the function u achieving the supremum in (1.24) satisfies

$$C \int_0^\ell u'v' dx = \int_0^\ell uv dx$$

13

for all v having square integrable first derivatives and vanish-
ing at 0 and ℓ. By integration by parts we then find

$$(1.25) \qquad \begin{cases} -u'' = \dfrac{1}{C}u, \ 0 < x < \ell \\ u(0) = u(\ell) = 0. \end{cases}$$

Thus $1/C$ is lowest eigenvalue of the eigenvalue problem (1.25),
and the optimal u in (1.24) (which achieves equality in (1.21))
is an associated eigenfunction.

B. Higher Dimensional Problems

The Vibrating Membrane

Consider the small, transverse vibration of a thin membrane
stretched over a bounded region Ω in the plane and fixed along
its edges $\Gamma = \partial\Omega$. The vertical displacement $u(x,y,t)$ of the
point (x,y) in Ω at time t satisfies

$$(1.26) \qquad \begin{cases} -\Delta u = -\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^2 u}{\partial y^2} = -\dfrac{\partial^2 u}{\partial t^2}, \ (x,y) \in \Omega, \ t > 0 \\ u(x,y,t) = 0, \ (x,y) \in \partial\Omega, \ t \geq 0. \end{cases}$$

As with the vibrating elastic bar or the vibrating string, if we
seek separated solutions of the form $u(x,y,t) = v(x,y)w(t)$, we
are led to the eigenvalue problem of finding λ and $v(x,y) \neq 0$
satisfying

$$(1.27) \qquad \begin{cases} -\Delta v = \lambda v, \ (x,y) \in \Omega \\ v(x,y) = 0, \ (x,y) \in \partial\Omega, \end{cases}$$

and for each eigenpair (λ, v) of (1.27), to the differential
equation

14

(1.28)
$$\frac{d^2w}{dt^2}(t) + \lambda w(t) = 0, \quad t \geq 0,$$

for $w(t)$ (cf. (1.16)).

It is known that (1.27) has an infinite sequence of eigen-values

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow +\infty$$

and corresponding eigenfunctions

$$v_1(x,y), \ v_2(x,y), \ \ldots \ .$$

The eigenfunctions are complete and orthonormal in $L_2(\Omega)$.
$a_j v_j(x,y) \sin \sqrt{\lambda_j}(t+\theta_j)$, $j = 1,2,\ldots$, are called eigenvibrations.
$\sqrt{\lambda_j}$ is the frequency and $v_j(x,y)$ is the shape of the jth
eigenvibration. All solutions of (1.26) can be obtained as a
superposition of eigenvibrations (cf. (1.18)). We note that if,
instead of fixing the membrane on Γ, we allowed it to move free-
ly in the vertical direction, then we whould have the Neumann boun-
dary condition $\frac{\partial u}{\partial n} = 0$, where $\frac{\partial}{\partial n}$ denotes the outer normal deri-
vative, instead of the Dirichlet condition $u = 0$. The approxima-
tion of the eigenpairs of a membrane is discussed in Subsection
10.B., 11.B., 12.A., and 12.B.

The Problem of Heat Conduction

Consider the problem of heat conduction in a body occupying a
region Ω in three-dimensional space. We suppose the temperature
distribution throughout Ω is known at time zero, the temperature
is held at zero on $\partial\Omega$ for all time, and that we want to deter-
mine the temperature $u(x,y,z,t)$ at the point $(x,y,z) \in \Omega$ at
time $t > 0$. From the fundamental law of heat conduction we know

15

that

$$\begin{cases} -\dfrac{\partial}{\partial x}(p(x,y,z)\dfrac{\partial u}{\partial x}) - \dfrac{\partial}{\partial y}(p(x,y,z)\dfrac{\partial u}{\partial y}) - \dfrac{\partial}{\partial z}(p(x,y,z)\dfrac{\partial u}{\partial z}) \\ \qquad\qquad\qquad = -r(x,y,z)\dfrac{\partial u}{\partial t}, \ (x,y,z) \in \Omega, \ t > 0 \\ u(x,y,z,t) = 0, \ (x,y,z) \in \partial\Omega, \ t \geq 0 \\ u(x,y,z,0) = f(x,y,z), \ (x,y,z) \in \Omega, \end{cases}$$

(1.29)

where

$f(x,y,z)$ = the temperature distribution at $t = 0$,

$p(x,y,z)$ = the thermal conductivity of the material at

$\qquad (x,y,z)$,

and

$r(x,y,z)$ = density of the material times the specific heat of

$\qquad$ the material.

If we seek separated solutions

$$u(x,y,z,t) = v(x,y,z)w(t)$$

of the differential equation and the boundary conditions in (1.29)
we are led to the eigenvalue problem

(1.30)
$$\begin{cases} -\dfrac{\partial}{\partial x}(p\dfrac{\partial v}{\partial x}) - \dfrac{\partial}{\partial y}(p\dfrac{\partial v}{\partial y}) - \dfrac{\partial}{\partial z}(p\dfrac{\partial v}{\partial z}) = \lambda rv, \ (x,y,z) \in \Omega \\ v(x,y,z) = 0, \ (x,y,z) \in \partial\Omega, \end{cases}$$

and for each eigenpair (λ,v) of (1.30) we are led to the equation

(1.31)
$$w' + \lambda w = 0, \ t > 0$$

for $w(t)$ (cf. (1.16) and (1.28)). (1.30) has eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow \infty$$

and eigenfunctions

$$v_1, v_2, \ldots$$

16

satisfying

$$\int_\Omega v_i \bar{v}_j r \ dx \ dy \ dz = \delta_{ij}.$$

Corresponding to each λ_j, from (1.31) we find $w(t) = w_j(t) = a_j e^{-\lambda_j t}$. Thus the separated solutions are given by

$$a_j v_j(x,y,z) e^{-\lambda_j t}, \ j = 1,2,\ldots,$$

and the solution of (1.29) is

$$(1.32) \qquad u(x,y,z,t) = \sum_{j=1}^\infty \left[\int_\Omega f \bar{v}_j r \ dx \ dy \ dz \right] v_j(x,y,z) e^{-\lambda_j t}$$

(cf. (1.18). We note that from (1.32) and the positivity of the eigenvalues, one can show that $\lim_{t\to\infty} u(x,y,z,t) = 0$ and that the rate at which the temperature u decays to zero is largely determined by λ_1.

The Vibration of an Elastic Solid

The vibration of an elastic solid Ω, the three-dimensional generalization of the elastic bar, is governed by the Navier-Lamé equations

$$(1.33) \quad \begin{cases} (\lambda+\mu)\dfrac{\partial\theta}{\partial x} + \mu\Delta u = -X + \rho\dfrac{\partial^2 u}{\partial t^2} \\[2mm] (\lambda+\mu)\dfrac{\partial\theta}{\partial y} + \mu\Delta v = -Y + \rho\dfrac{\partial^2 v}{\partial t^2} \\[2mm] (\lambda+\mu)\dfrac{\partial\theta}{\partial z} + \mu\Delta w = -Z + \rho\dfrac{\partial^2 w}{\partial t^2}, \ (x,y,z) \in \Omega, \ t > 0, \end{cases}$$

where $u(x,y,z,t)$, $v(x,y,z,t)$, and $w(x,y,z,t)$ are the $x,y,$ and z-components of the displacement of the point $(x,y,z) \in \Omega$

17

at time t, $\theta = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$, X, Y, and Z are the components of the external force per unit volume acting at (x, y, z), $\lambda > 0$, and $\mu > 0$ are the Lamé elastic constants, and ρ is the density of the material.

As in the case of the bar, boundary conditions of various types may be prescribed. For example, the Dirichlet boundary conditions prescribe the values of u, v, and w on $\Gamma = \partial\Omega$. Neumann conditions are more complicated. Let n be the unit outer normal to Γ, let n_x, n_y, and n_z be the x, y, and z-components of n, and let

$$\frac{\partial}{\partial n} = n_x \frac{\partial}{\partial x} + n_y \frac{\partial}{\partial y} + n_z \frac{\partial}{\partial z}$$

be the outer normal derivative. Then define

(1.34a) $\qquad X_n = \lambda\theta n_x + \mu\frac{\partial u}{\partial n} + \mu[\frac{\partial u}{\partial x} n_x + \frac{\partial v}{\partial x} n_y + \frac{\partial w}{\partial x} n_z]$

(1.34b) $\qquad Y_n = \lambda\theta n_y + \mu\frac{\partial v}{\partial n} + \mu[\frac{\partial u}{\partial y} n_x + \frac{\partial v}{\partial y} n_y + \frac{\partial w}{\partial y} n_z]$

(1.34c) $\qquad Z_n = \lambda\theta n_z + \mu\frac{\partial w}{\partial n} + \mu[\frac{\partial u}{\partial z} n_x + \frac{\partial v}{\partial z} n_y + \frac{\partial w}{\partial z} n_z].$

The Neumann conditions then consist in prescribing X_n, Y_n, and Z_n on the boundary. One can also mix the boundary conditions in various ways, e.g., impose Dirichlet conditions on one part of the boundary and Neumann conditions on the remainder of the boundary or prescribe X_n, Y_n, and w on Γ.

The eigenvalue problem associated with (1.33) is given by

(1.35) $\qquad \begin{cases} -(\lambda+\mu)\dfrac{\partial\theta}{\partial x} - \mu\Delta u = \omega\rho u \\[2mm] -(\lambda+\mu)\dfrac{\partial\theta}{\partial y} - \mu\Delta v = \omega\rho v \\[2mm] -(\lambda+\mu)\dfrac{\partial\theta}{\partial z} - \mu\Delta w = \omega\rho w, \quad (x, y, z) \in \Omega, \end{cases}$

18

where we have denoted the eigenvalue parameter by ω (to avoid confusion with the Lamé constants μ and λ), and where here u,v,w, and θ denote functions of x,y, and z only, i.e., the separation of variables has been written as $u(x,y,z,t) = u(x,y,z)$ $T(t)$, etc. For boundary conditions we can consider any of those mentioned above. If we consider Dirichlet conditions $(u = v = w = 0$ on $\Gamma)$ we refer to the clamped solid and if we consider Neumann conditions $(X_n = Y_n = Z_n = 0$ on $\Gamma)$ we refer to the free solid.

The approximation of the eigenvalues of the free L-shaped panel (a two dimensional analogue of the elastic solid) is treated in detail in Subsection 10.A.

The Steklov Eigenvalue Problem

The Steklov eigenvalues of the differential operator $-\Delta + I$ are those numbers λ such that for some nonzero u,

$$\begin{cases} -\Delta u + u = 0 & \text{in } \Omega \\ \dfrac{\partial u}{\partial n} = \lambda u & \text{on } \Gamma = \partial\Omega. \end{cases}$$

Problems of this type, in which the eigenvalue parameter appears in the boundary condition, arise in a number of applications (cf. (1.9b) and (1.9c)).

The Problem of Stability of a Nonlinear Problem

Consider the quasilinear parabolic problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u + u\dfrac{\partial u}{\partial x} = 0, & (x,y) \in \Omega, \; t > 0 \\ u(x,y,t) = \varphi(x,y), & (x,y) \in \partial\Omega, \; t > 0. \end{cases}$$

Suppose $\tilde{u}(x,y)$ is a stationary solution, i.e., suppose

$$\begin{cases} -\Delta\tilde{u} + \tilde{u}\dfrac{\partial\tilde{u}}{\partial x} = 0, & (x,y) \in \Omega, \\[2mm] \tilde{u}(x,y) = \varphi(x,y), & (x,y) \in \partial\Omega. \end{cases}$$

Then we consider a nearby time-dependent solution

$$u(x,y,z,t) = \tilde{u}(x,y) + w(x,y,t)$$

and ask whether $\tilde{u}$ is a <u>stable stationary solutions</u>, i.e., whether

$$\lim_{t\to\infty} u(x,y,t) = \tilde{u}(x,y)$$

or, equivalently,

$$\lim_{t\to\infty} w(x,y,t) = 0.$$

We easily see that w satisfies

(1.36)
$$\begin{cases} \dfrac{\partial w}{\partial t} + Lw + Nw = 0, & (x,y) \in \Omega, \ t > 0 \\[2mm] w = 0, & (x,y) \in \partial\Omega, \end{cases}$$

where

$$Lw = -\Delta w + \tilde{u}\frac{\partial w}{\partial x} + \frac{\partial\tilde{u}}{\partial x}w$$

and

$$Nw = w\frac{\partial w}{\partial x}.$$

Conditions ensuring $w \to 0$ as $t \to \infty$ can be given in terms of the eigenvalues of

$$\begin{cases} Lw = \lambda w & \text{in } \Omega \\[2mm] w = 0 & \text{on } \partial\Omega. \end{cases}$$

In fact, if all the eigenvalues of this problem have positive real parts, then $\tilde{u}$ is asymptotically stable in the L_2 norm, i.e., there is a constant $\delta > 0$ such that if

20

$$\|w(\cdot,\cdot,0)\|_{L_2(\Omega)} < \delta,$$

then

$$\|w(\cdot,\cdot,t)\|_{L_2(\Omega)} \to 0 \quad \text{as} \quad t \to \infty.$$

If the term N in (1.36) is neglected, then this result is simi-
lar to that mentioned at the end of the discussion of heat conduc-
tion. Note that L is a nonselfadjoint operator and its eigen-
values will, in general, be complex (cf. Section 3). For further
detail on this type of stability results see Prodi [1962].

Section 2.   Sobolev Spaces

The natural setting for a discussion of eigenvalue problems and their approximation is the theory of linear operators on a Hilbert space.   In this section we will sketch the definitions and basic properties of the function spaces we will make use of. These are mainly the Sobolev and Besov spaces.

Let Ω be a bounded open subset of R^n and denote by $x = (x_1,\ldots,x_n)$ a point in R^n.   For each integer $m \geq 0$, the real (complex) Sobolev space $H^m(\Omega)$ is defined by

$$(2.1) \qquad H^m = H^m(\Omega) = \{u : \partial^\alpha u \in L_2(\Omega) \ \forall \ |\alpha| \leq m\},$$

where $L_2(\Omega)$ denotes the usual space of real (complex) valued square-integrable functions on Ω equipped with the inner product

$$(2.2) \qquad (u,v) = (u,v)_{L_2(\Omega)} = \int_\Omega u\bar{v}dx$$

and norm

$$(2.3) \qquad \|u\| = \|u\|_{L_2(\Omega)} = (\int_\Omega |u|^2 dx)^{1/2}.$$

On $H^m(\Omega)$ we have the inner product

$$(2.4) \qquad ((u,v))_m = ((u,v))_{m,\Omega} = \sum_{|\alpha| \leq m} \int_\Omega \partial^\alpha u \overline{\partial^\alpha v} \ dx$$

and norm

$$(2.5) \qquad \|u\|_{H^m(\Omega)} = \|u\|_m = \|u\|_{m,\Omega} = (\sum_{|\alpha| \leq m} \int_\Omega |\partial^\alpha u|^2 dx)^{1/2}.$$

With this inner product, $H^m(\Omega)$ is a Hilbert space.   Here $\alpha = (\alpha_1,\ldots,\alpha_n)$, with the α_i a nonnegative integer, $|\alpha| = \sum_i \alpha_i$, and $\partial^\alpha u = \partial^{|\alpha|} u/\partial x_1^{\alpha_1},\ldots,\partial x_n^{\alpha_n}$.   We also have the semi-inner product

$$(2.6) \qquad (u,v)_{H^m(\Omega)} = (u,v)_m = (u,v)_{m,\Omega} = \sum_{|\alpha|=m} \int_\Omega \partial^\alpha u \overline{\partial^\alpha v} \, dx$$

and semi-norm

$$(2.7) \qquad |u|_{H^m(\Omega)} = |u|_m = |u|_{m,\Omega} = \left(\sum_{|\alpha|=m} \int_\Omega |\partial^\alpha u|^2 dx \right)^{1/2}.$$

It is immediate that $H^0(\Omega) = L_2(\Omega)$ and $\|u\|_{0,\Omega} = |u|_{0,\Omega} = \|u\|_{L_2(\Omega)}$. If $\Gamma = \partial\Omega$ is Lipschitz continuous, then $C^m(\bar{\Omega})$ is dense in $H^m(\Omega)$. (Γ is called Lipschitz continuous if it can be locally represented by a Lipschitz continuous function; see Necas [1967] for further details.)

$H_0^1(\Omega)$ is defined as the closure in $H^1(\Omega)$ of $C_0^\infty(\Omega)$, the space of infinitely differentiable functions on Ω which vanish near Γ. The Poincaré inequality, which states that

$$(2.8) \qquad |u|_{0,\Omega} \le C |u|_{1,\Omega}, \quad \forall \, u \in H_0^1(\Omega),$$

shows that $|\cdot|_{1,\Omega}$ is a norm on $H_0^1(\Omega)$. $H_0^m(\Omega)$ is the closure in $H^m(\Omega)$ of $C_0^\infty(\Omega)$.

If Γ is Lipschitz continuous, then we can define the space $L_2(\Gamma)$, which consists of functions u defined on Γ for which $\|u\|_{L_2(\Gamma)} = \left(\int_\Gamma |u|^2 ds \right)^{1/2} < \infty$, where ds denotes the surface area. $L_2(\Gamma)$ is a Hilbert space with inner product $(u,v)_{L_2(\Gamma)} = \int_\Gamma u\bar{v} ds$. It is also known that a function $u \in H^1(\Omega)$ has a well-defined restriction to Γ, denoted by $tr \, u$, in the sense of trace; $u = tr \, u$ satisfies

$$(2.9) \qquad \|u\|_{L_2(\Gamma)} \le C \|u\|_{1,\Omega}, \quad \forall \, u \in H^1(\Omega),$$

and

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma \text{ in the sense of trace}\}.$$

Furthermore, a function $u \in C^1(\bar{\Omega})$ is in $H_0^1(\Omega)$ if an only if $u = 0$ for all $x \in \Gamma$. We note that if Γ is Lipschitz continuous, then the normal vector n is defined almost everywhere on Γ. The outer normal derivative $\frac{\partial u}{\partial n}$ is defined for $u \in H^2(\Omega)$.

$$H_0^2(\Omega) = \{u \in H^2(\Omega) : u = \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma\}.$$

We shall occasionally make use of the vector valued Sobolev spaces $\vec{H}^m(\Omega)$ which are defined by

$$(2.10) \quad \vec{H}^m(\Omega) = \{(u_1(x),\ldots,u_k(x)) : u_j(x) \in H^m(\Omega), \, j = 1,\ldots,k\}$$

and

$$(2.11) \qquad \|u\|_{\vec{H}^m(\Omega)}^2 = \|u_1\|_{m,\Omega}^2 + \ldots + \|u_k\|_{m,\Omega}^2.$$

In the study of eigenvalue problems, central use will be made of Rellich's theorem (cf. Agmon [1965]), which states that every bounded sequence in $H^m(\Omega)$ has a subsequence which converges in $H^j(\Omega)$ if $j < m$, provided Ω is a bounded open set in R^n with a Lipschitz continuous boundary.

So far we have defined the Sobolev space $H^m(\Omega)$ only for m an integer. We will sometimes use $H^m(\Omega)$, for m fractional and also the Besov spaces, so we now turn to their definition, using the K-method.

For $u \in H^m(\Omega)$ and $0 < t < \infty$ set

$$(2.12) \qquad K(u,t) = \inf_{\substack{v \in H^m, w \in H^{m+1} \\ v+w=u}} \{\|v\|_{m,\Omega} + t\|w\|_{m+1,\Omega}\}.$$

Then for $m < k < m+1$ define

24

$$(2.13) \qquad \|u\|_{H^k(\Omega)} = \|u\|_k = \|u\|_{k,\Omega} = \left(\int_0^{\infty} [t^{-\theta} K(t,u)]^2 \, \frac{dt}{t} \right)^{1/2}$$

and

$$(2.14) \qquad \|u\|_{\hat{H}^k(\Omega)} = \sup_{0 < t < \infty} \{ t^{-\theta} K(u,t) \},$$

where $\theta = k-m$. The space

$$(2.15) \qquad H^k(\Omega) \equiv \{ u \in H^m(\Omega) : \|u\|_{H^k(\Omega)} < \infty \}$$

is the Sobolev space with fractional order k and

$$(2.16) \qquad \hat{H}^k(\Omega) \equiv \{ u \in H^m(\Omega) : \|u\|_{\hat{H}^k(\Omega)} < \infty \}$$

is a Besov space, the one often denoted by $B^k_{2,\infty}$.

In order to fix these ideas and to obtain a fact we will use in the sequel (cf. Subsections 10.A. and 10.B.), we now consider the function

$$u = r^{\alpha}, \text{ for } (r,\theta) \in S = \{ (r,\theta) : 0 < r < 1, \ 0 \leq \theta \leq \theta_0 \leq 2\pi \},$$

where $-1 < \alpha < 0$, (r,θ) being polar coordinates, and prove that $u \in \hat{H}^{\alpha+1}(S)$.

Theorem 2.1. For $-1 < \alpha < 0$, we have
$$u = r^{\alpha} \in \hat{H}^{1+\alpha}(S).$$

Proof. Let $\varphi(x)$, $0 < x < \infty$, be a function having derivatives of all orders and satisfying

$$\varphi(x) = 0 \text{ for } 0 < x < 1/2,$$

$$\varphi(x) = 1 \text{ for } 1 < x < \infty.$$

For $0 < \delta \leq 1$, define

$$v = [1 - \varphi(\tfrac{r}{\delta})] u,$$

25

$$w = \varphi\left(\frac{r}{\delta}\right)u.$$

Then we obviously have $u = v+w$. Now

$$\|v\|^2_{H^0(S)} \leq \theta_0 \int_0^\delta r^{2\alpha+1}dr$$

$$= \frac{\delta^{2\alpha+2}\theta_0}{2\alpha+2}$$

$$\leq C\delta^{2\alpha+2}$$

and

$$\|w\|^2_{H^1(S)} = \int_S \left[|w|^2 + |\frac{\partial w}{\partial x_1}|^2 + |\frac{\partial w}{\partial x_2}|^2\right]dx_1 dx_2$$

$$= \int_0^{\theta_0}\int_0^1 \left[|w|^2 + |\frac{\partial w}{\partial r}|^2 + r^{-2}|\frac{\partial w}{\partial \theta}|^2\right]r\, dr d\theta$$

$$= \theta_0 \int_0^1 \left[|w|^2 + |\frac{\partial w}{\partial r}|^2\right]r\, dr$$

$$\leq C[\int_{\delta/2}^1 r^{2\alpha+1}dx + \alpha^2\int_{\delta/2}^1 r^{2\alpha-1}dr$$

$$+ \delta^{-2}\int_{\delta/2}^\delta r^{2\alpha+1}dr]$$

$$\leq C\delta^{2\alpha},$$

with C independent of δ. Hence

$$K(u,t) \leq C[\delta^{\alpha+1} + t\delta^\alpha]$$

and thus

$$t^{-(\alpha+1)}K(u,t) \leq C[\delta^{\alpha+1}t^{-(\alpha+1)} + t^{1-(\alpha+1)}\delta^\alpha].$$

If $0 < t < 1$, let $\delta = t$ to get

$$t^{-(a+1)}K(u,t) \le 2C$$

and hence

$$\sup_{0<t<1} \{t^{-(a+1)}K(u,t)\} \le 2C.$$

If $t \ge 1$, we obviously have

$$K(u,t) = |u|_{H^0(\Omega)} \le C'$$

and hence

$$\sup_{1\le t<\infty} \{t^{-(a+1)}K(u,t)\} \le C'.$$

Therefore

$$|u|_{\hat{H}^{(1+a)}(S)} = \sup_{0<t<\infty} \{t^{-(1+a)}K(u,t)\} \le C'' < \infty$$

and hence $u \in \hat{H}^{1+a}(S)$, as was to be proved.

In a similar way, one can also prove that $r^a \in \hat{H}^{1+a}(S)$ for $a > 0$, not an integer. Finally we note that $r^a \in H^{1+a}(S)$, but $r^a \in H^{1+a-\varepsilon}(S)$ for any $\varepsilon > 0$.

For a complete discussion of the Sobolev and Besov spaces we refer to Adams [1975], Nečas [1967], and Butzer and Barens [1967].

Remark 2.1. The definition of the Sobolev spaces with fractional index m has a very simple interpretation. For u to be in $\hat{H}^{1+a}(S)$ means that for any $0 < t < \infty$, u can be split into the sum of a smooth function and a nonsmooth function in a natural way. We have employed this natural splitting in the proof of Theorem 2.1 and we will use it in the sequel.

So far we have considered only one special family of Sobolev spaces or Sobolev-type spaces. Several other families are impor-

·tant in various situations. For example, if $\Omega \subset R^2$ with $0 \in \partial\Omega$, and if $0 < \beta < 1$ and $m \geq \ell \geq 1$, we can define

(2.17)
$$H_\beta^{m,\ell}(\Omega) = \{u \in H^{\ell-1}(\Omega) : (\partial^\alpha u)r^{\beta+|\alpha|-\ell} \in L_2(\Omega) \quad \text{for} \quad \ell \leq |\alpha| \leq m\}$$

and

$$(2.18) \quad \|u\|_{H_\beta^{m,\ell}(\Omega)}^2 = \|u\|_{H^{\ell-1}(\Omega)}^2 + \sum_{|\alpha|=\ell}^{m} \|(\partial^\alpha u)r^{\beta+|\alpha|-\ell}\|_{L_2(\Omega)}^2,$$

where $r = (x_1^2 + x_2^2)^{1/2}$. Spaces of this kind are called weighted Sobolev spaces. For more details we refer to Kufner [1985]. Consider the function $u = r^\gamma$, with $0 < \gamma < 1$. One can show that $u \in H_\beta^{m,\ell}(\Omega)$, where $\Omega = \{(r,\theta) : 0 < r < 1\}$, for $\beta > 1-\gamma$, $m = 2$, and $\ell = 2$. In fact, since $|\partial^\alpha u| \leq C(\alpha)r^{\gamma-|\alpha|}$, we have $|\partial^\alpha u|r^{\beta+|\alpha|-2} \leq r^{\gamma+\beta-2}$, and we see that $u \in H_\beta^{m,\ell}(\Omega)$ for m, ℓ, and β as given.

We will also have occasion to use countably normed spaces constructed from Sobolev spaces. For example, consider the space

$$(2.19) \quad \mathcal{B}_\beta^2(\Omega) = \{u \in H_\beta^{2,2}(\Omega) : \|(\partial^\alpha u)r^{|\alpha|-2+\beta}\|_{L_2(\Omega)} \leq Cd^{|\alpha|}\alpha!$$

for $|\alpha| > 2$, with C and d independent of $\alpha\}$.

It is easy to see that all functions $u \in \mathcal{B}_\beta^2(\Omega)$ are analytic in $\bar\Omega - \{0\}$. The function r^γ considered above belongs to $\mathcal{B}_\beta^2(\Omega)$ for $\beta > 1-\gamma$. We have here only considered weights with respect to the origin. More generally, one can consider weights with respect to the vertices of domains with piecewise smooth boundaries. An important reason for introducing these spaces is to characterize the solution (eigenfunctions) of a problem as precisely as possible

28

by embedding it (them) in as small a space as possible. There are other classes of function spaces that are important in various contexts, but we will not go further in this direction.

Remark 2.2. We have followed the usual custom of using the same notation for real and complex function spaces. It will be clear from the context which version we are using. See Remark 4.1.

Section 3. Variational Formulation of Eigenvalue Problems

In Section 1 the eigenvalue problems were stated in classical form, i.e., we were seeking an eigenvalue λ and a corresponding nonzero eigenfunction $u(x)$ such that the eigenvalue equation and boundary condition were satisfied in the classical pointwise sense. These problems can alternately be given variational formulations. Since finite element approximation methods are most naturally defined in terms of variational formulations we now briefly indicate how eigenvalue problems can be cast in variational form. We will do this by discussing 2nd order elliptic eigenvalue problems in two dimensions in some detail. We begin by describing this type of problem.

Consider the problem:

Seek a real or complex number λ and a nonzero real or complex valued function $u(x)$ satisfying

$$(3.1) \qquad \begin{cases} (Lu)(x) = \lambda(Mu)(x), \ x \in \Omega \\ (Bu)(x) = 0, \ x \in \Gamma = \partial\Omega, \end{cases}$$

where Ω is a bounded, open, connected set in R^2, and

$$(3.2) \quad Lu(x) = -\sum_{i,j=1}^{2} \partial_j(a_{ij}(x)\partial_i u) + \sum_{i=1}^{2} b_i(x)\partial_i u + c(x)u, \quad (\partial_i = \tfrac{\partial}{\partial x_i})$$

where $a_{ij}(x) = a_{ji}(x)$, $b_i(x)$, and $c(x)$ are given real or complex functions on Ω,

$$(3.3) \qquad\qquad Mu(x) = d(x)u(x),$$

where $d(x)$ is a given real function which is bounded below by a positive constant on Ω, and

$$(3.4) \qquad (Bu)(x) = \begin{cases} u(x) \\ \text{or} \\ -\displaystyle\sum_{i,j=1}^{2} a_{ij}n_j\partial_i u, \end{cases}$$

where $n(x) = (n_1, n_2)$ is the exterior unit normal to $\Gamma = \partial\Omega$ at

x. L is assumed to be uniformly strongly elliptic in Ω, i.e.,

there is a positive constant a_0 such that

$$(3.5) \quad \text{Re} \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \geq a_0 \sum_{i=1}^{2} \xi_i^2, \ \forall \ x \in \Omega \quad \text{and} \quad \forall (\xi_1, \xi_2) \in R^2.$$

In addition, $a_{ij}, b_i, c,$ and d are assumed to be bounded and

measurable. (A portion of the theory of eigenvalue problems can

be developed under the more general hypothesis that $d(x)$ is

merely assumed to be a bounded, measurable, complex function, but

we will not pursue this direction.)

(λ, u) is called an eigenpair of the 2nd order differential

operator L (relative to the 0th order differential operator

M). If $Bu = u$, the boundary condition $Bu = 0$ is the Dirichlet

condition, and if $Bu = -\displaystyle\sum_{i,j=1}^{2} a_{ij}n_j\partial_i u = \dfrac{\partial u}{\partial \nu}$ = the conormal deriva-

tive of u, then $Bu = 0$ yields the Neumann condition.

It is immediate that all of the examples discussed in Sec-

tion 1 — except the Steklov-type eigenvalue problems and the prob-

lem of the vibration of an elastic solid — are of the form (3.1)

or its one or higher dimensional analogues. In any case, our dis-

cussion of approximation methods will be in terms of an abstract

framework that will cover all the examples.

Let

$$(3.6) \qquad L^*v(x) = -\sum_{i,j=1}^{2} \partial_i(\overline{a_{ij}}\partial_j v) - \sum_{i=1}^{2}{}_i\partial_i(\bar{b}_i v) + \bar{c}v$$

and

$$(3.7) \qquad \frac{\partial v}{\partial \nu}_* = -\sum_{i,j=1}^{2} \overline{a_{ij}}n_i\partial_j v - \sum_{i=1}^{2}{}_i\bar{b}_i n_i v.$$

L^* is called the formal adjoint of L. It is an immediate consequence of the divergence theorem that

$$(3.8) \qquad \int_\Omega Lu\bar{v}dx = \int_\Omega \left(\sum_{i,j=1}^{2} a_{ij}\partial_i u\overline{\partial_j v} + \sum_{i=1}^{2} b_i\partial_i u\bar{v} + cu\bar{v}\right)dx + \int_\Gamma \frac{\partial u}{\partial \nu}\bar{v}ds$$

$$= \int_\Omega u\overline{L^*v}dx + \int_\Gamma \frac{\partial u}{\partial \nu}\bar{v}dx - \int_\Gamma u\overline{\frac{\partial v}{\partial \nu}}_*ds$$

for all smooth functions u and v. Hence we have

$$(3.9) \qquad \int_\Omega Lu\bar{v}\ dx = \int_\Omega u\overline{L^*v}dx$$

if either $u = v = 0$ on Γ or $\dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu}_* = 0$ on Γ.

If a_{ij} and c are real and $b_i = 0$, then $L^* = L$ and $\dfrac{\partial}{\partial \nu} = \dfrac{\partial}{\partial \nu_*}$. In this case we say L,M,B or, more briefly, L is for-mally selfadjoint, and we have

$$(3.10) \qquad \int_\Omega d(\tfrac{1}{d}L)u\bar{v}\ dx = \int_\Omega du\overline{(\tfrac{1}{d}L)v}\ dx$$

if either $u = v = 0$ on Γ or $\dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = 0$ on Γ. All of the examples treated in Section 1 are formally selfadjoint except the operator arising in the stability analysis of the nonlinear initial-boundary value problem.

Now we turn to the derivation of a variational formulation for (3.1). Suppose $(\lambda, u(x))$ satisfies (3.1) in the classical sense, i.e., the differential equation and the boundary condition hold pointwise, and consider first the case of the Dirichlet boundary condition: $u(x) = 0$ for $x \in \Gamma$. We assume Ω is a bounded open set in R^2 with Lipschitz continuous boundary Γ. Then, multiplying (3.1) by v, integrating over Ω, and using (3.3) and (3.8) we find that

$$(3.14) \quad \lambda b(u,v) \equiv \lambda \int_{\Omega} du\bar{v}dx = \int_{\Omega} Lu\bar{v}dx$$

$$= \int_{\Omega} \left[\sum_{i,j=1}^{2} a_{ij}\partial_i u\overline{\partial_j v} + \sum_{i=1}^{2} b_i \partial_i u\bar{v} + cu\bar{v} \right] dx$$

$$+ \int_{\Gamma} \frac{\partial u}{\partial \nu} \bar{v}ds$$

$$= \int_{\Omega} \left[\sum_{i,j=1}^{2} a_{ij}\partial_i u\overline{\partial_j v} + \sum_{i=1}^{2} b_i \partial_i u\bar{v} + cu\bar{v} \right] dx$$

$$\equiv a(u,v), \quad \text{for all } v \in C^1(\bar{\Omega})$$
$$\text{that vanish on } \Gamma.$$

$a(u,v)$ and $b(u,v)$, as defined in (3.14), are bilinear forms (sometimes referred to as sesquilinear forms in the complex case) in u and v. They are clearly defined for $u,v \in C^1(\bar{\Omega})$ and, in fact, $a(u,v)$ is defined for $u,v \in H^1(\Omega)$ and $b(u,v)$ for $u,v \in H^0(\Omega) = L_2(\Omega)$. Furthermore, using the fact that a_{ij}, b_i, c, and d are bounded, it follows from Schwarz's inequality that a is bounded on $H^1(\Omega)$ and b is bounded on $H^0(\Omega)$, i.e., that

$$(3.15) \qquad |a(u,v)| \leq C_1 \|u\|_{1,\Omega} \|v\|_{1,\Omega}, \quad \forall\ u,v \in H^1(\Omega),$$

$$(3.16) \qquad |b(u,v)| \leq C_2 \|u\|_{0,\Omega} \|v\|_{0,\Omega}, \quad \forall\ u,v \in H^0(\Omega).$$

We note one further property of the form $a(u,v)$:

$$(3.17) \quad \text{Re}\ a(u,u) \geq \begin{cases} \dfrac{a_0}{2}\, \|u\|^2_{1,\Omega}, \quad \forall\ u \in H^1(\Omega), \quad \text{provided} \\[2mm] \quad \text{Re}\ c(x) \geq \dfrac{a_0}{2} + \dfrac{b^2}{2a_0}, \quad \text{for all}\ x \in \Omega, \\[2mm] \quad \text{where}\ b = \max_{\substack{x \in \Omega \\ i=1,2}} |b_i(x)|, \\[4mm] a_0 |u|^2_{1,\Omega} \geq C\|u\|^2_{1,\Omega}, \quad \forall\ u \in H^1_0(\Omega), \quad \text{provided} \\[2mm] \quad b_i(x) = 0,\ i = 1,2,\ \text{Re}\ c(x) \geq 0. \end{cases}$$

a_0 here is the ellipticity constant in (3.5); the result follows directly from (3.5).

Since the eigenfunction u vanishes on Γ, $u \in H^1_0(\Omega)$. Thus, using (3.15), (3.16), and the fact that $\{v \in C^1(\bar{\Omega}) : v = 0$ on $\Gamma\}$ is dense in $H^1(\Omega)$, it follows from (3.14) that the eigenpair (λ,u) satisfies

$$(3.18) \qquad \begin{cases} u \in H^1_0(\Omega), \quad u \neq 0 \\[2mm] a(u,v) = \lambda\, b(u,v), \quad \forall\ v \in H^1_0(\Omega). \end{cases}$$

(3.18) is called a variational formulation of (3.1). We have shown that if (λ,u) is an eigenpair in the classical sense then it is an eigenpair in the variational sense. We now show that the converse is true, provided $\Gamma, a_{ij}, b_i, c,$ and d are sufficiently smooth.

Suppose (λ,u) satisfies (3.18) and suppose in addition Ω

34

is a bounded open set with Lipschitz continuous boundary Γ and that $u \in C^2(\bar{\Omega})$. Then from the equation in (3.18) and from (3.8) we have

$$(3.19) \qquad \int_\Omega Lu\bar{v}dx = a(u,v) + \int_\Omega \frac{\partial u}{\partial \nu}\bar{v}ds$$

$$= a(u,v)$$

$$= \lambda b(u,v)$$

$$= \lambda \int_\Omega du\bar{v}dx, \ \forall \ v \in C^1(\bar{\Omega}) \ \text{that vanishes on} \ \Gamma.$$

Since $\{v \in C^1(\bar{\Omega}) : v = 0 \text{ on } \Gamma\}$ is dense in $L_2(\Omega)$ we see from (3.19) that

$$Lu(x) = \lambda Mu(x), \ x \in \Omega.$$

Also, since Γ is Lipschitz continuous and $u \in C^2(\bar{\Omega}) \cap H_0^1(\Omega)$ we know that $u(x) = 0$ for all $x \in \Gamma$. Thus we see that (λ, u) is an eigenpair in the classical sense.

We next present conditions that guarantee that $u \in C^2(\bar{\Omega})$. From (3.18) we see that u is a weak solution of the boundary value (source) problem

$$\begin{cases} Lu = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma, \end{cases}$$

where $f = \lambda du$. Using standard regularity results for elliptic equations we find that $u \in C^2(\bar{\Omega})$ provided $\Gamma, a_{ij}, b_i, c,$ and d are sufficiently smooth. In the two-dimensional case we are discussing it is sufficient to assume

· Γ is of class C^4,

· $a_{ij}, b_i \in C^3(\bar{\Omega})$, and

35

· $c, d \in C^2(\bar{\Omega})$.

In the general n-dimensional case it is sufficient to assume

· Γ is of class C^k,

· $a_{ij}, b_i \in C^{k-1}(\bar{\Omega})$, and

· $c, d \in C^{k-2}(\bar{\Omega})$, where $k = [n/2] + 3$.

For these results we refer to Agmon [1965, Theorems 3.9 and 9.8].

Eigenvalue problems on domains with corners arise in many applications but are not covered by the above results because of the requirement that Γ be smooth. Nevertheless, when Ω has corners, analogous results in a generalized form involving weighted Sobolev spaces can be proved for problems with smooth coefficients (see Grisvard [1985] and Babuška and Guo [1987]). Furthermore these results provide information on the behavior of u near the corners that is important in assessing the accuracy of eigenvalue approximations. This matter will be taken up in Section 10. We now briefly outline the extent to which the eigenpair (λ, u) of (3.18) satisfies (3.1) in the classical sense in the case in which Ω is a polygon and $L = -\Delta$ and $d(x) = 1$. From regularity results for elliptic equations we conclude that $u \in C^\infty(\bar{\Omega} - \{vertices$ of $\Omega\})$. Thus we see that $Lu(x) = \lambda Mu(x)$ for all $x \in \Omega$ and $u(x) = 0$ for $x \in \Gamma - \{vertices$ of $\Omega\}$. u fails, however, to be an eigenfunction in the classical sense in that $u \in C^2$ at any vertex of Ω with interior angle larger than π.

Under the hypothesis sketched above, the classical and variational formulations of (3.1) are equivalent. For the remainder of this article, we will take the point of view that our eigenvalue problems are given in variational form. Thus we will consider

36

problems of the form (3.18), or problems that are generalizations
of the form (3.18); see Section 8.

Consider now the case of the Neumann boundary condition:
$\frac{\partial u}{\partial \nu}(x) = 0$ for $x \in \Gamma$. Suppose (λ, u) satisfies (3.1) in the clas-
sical sense. Then, using (3.8) we find

$$\lambda b(u,v) = a(u,v) + \int_\Gamma \frac{\partial u}{\partial \nu} \bar{v} ds$$

$$= a(u,v), \quad \text{for all} \quad v \in C^1(\bar{\Omega}),$$

and thus, using the fact that $C^1(\bar{\Omega})$ is dense in $H^1(\Omega)$, we see
that (λ, u) satisfies

(3.20)
$$\begin{cases} u \in H^1(\Omega), \quad u \neq 0 \\ a(u,v) = \lambda b(u,v) \; \forall \; v \in H^1(\Omega). \end{cases}$$

(3.20) is a variational form for (3.1) with the Neumann condition.
Now suppose (λ, u) satisfies (3.20) and assume $u \in C^2(\bar{\Omega})$. From
(3.20) and (3.8) we obtain

(3.21)
$$\int_\Omega Lu\bar{v}dx = a(u,v) + \int_\Gamma \frac{\partial u}{\partial \nu} \bar{v}dx$$

$$= \lambda b(u,v) + \int_\Gamma \frac{\partial u}{\partial \nu} \bar{v}ds$$

$$= \lambda \int_\Omega du\bar{v}dx + \int_\Gamma \frac{\partial u}{\partial \nu} \bar{v}ds, \; \forall \; v \in C^1(\bar{\Omega}).$$

Taking $v \in C^1(\bar{\Omega})$ which vanish on Γ we find that

$$Lu(x) = \lambda u(x), \; \forall \; x \in \Omega.$$

Thus (3.21) reduces to

$$\int_{\Gamma} \frac{\partial u}{\partial \nu}\, \bar{v}\, ds = 0, \quad \forall\ v \in C^1(\bar{\Omega}),$$

which implies that $\frac{\partial u}{\partial \nu} = 0$ on Γ. Thus we have shown that (λ, u) satisfies (3.1) in the classical sense. As with the Dirichlet condition, the analysis is valid under appropriate smoothness assumptions on $\Gamma, a_{ij}, b_i, c,$ and d. We will not state these in detail.

Note that the Neumann boundary condition is not explicitly stated in (3.20). It is, however, implicitly contained in (3.20). We refer to the Neumann condition as a <u>natural</u> <u>boundary</u> <u>condition</u>, in contrast to the Dirichlet condition which is referred to as an <u>essential</u> <u>boundary</u> <u>condition</u>, and which is explicitly contained in the variational formulation (3.18). The fact that the Neumann condition is natural has important implications for the approximation of eigenvalues; see Remark 10.5.

In summary, for (3.1) we get one of the following forms:

<u>Problem 1: Dirichlet boundary condition</u>

Seek $\lambda, u \neq 0$ satisfying

$$\begin{cases} u \in H^1_0(\Omega) \\ a(u,v) = \lambda b(u,v), \quad \forall\ v \in H^1_0(\Omega) \end{cases}$$

<u>Problem 2: Neumann boundary condition</u>

Seek $\lambda, u \neq 0$ satisfying

$$\begin{cases} u \in H^1(\Omega) \\ a(u,v) = \lambda b(u,v), \quad \forall\ v \in H^1(\Omega) \end{cases}$$

We will sometimes refer to (λ, u) as an eigenpair of the form a

relative to the form b. Regarding the forms a and b we assume
(3.15) - (3.17) hold.

In a similar way, many other problems — including all of the
examples discussed in Section 1 — can be given variational formu-
lations. This is done for a number of problems in Chapter III.
We mention in particular the eigenvalue problems corresponding to
the vibration of a free L-shaped panel (a two dimentional ana-
logue of the elastic solid).

Finally we wish to make one further point regarding varia-
tional formulations of eigenvalue problems, namely, that a given
eigenvalue problem can often be given a variety of different
variational formulations and that some of these may lead to more
effective finite element methods than others. We illustrate the
possibility of various variational formulations by considering the
simple model problem

$$(3.22) \qquad \begin{cases} -(a(x)u')' + cu = \lambda u, \ 0 < x < 1 \\ u(0) = u(1) = 0. \end{cases}$$

This has already been cast into the variational form

$$(3.23) \qquad \begin{cases} \text{Seek } \lambda, u \neq 0 \text{ satisfying} \\ u \in H_0^1(0,1) \\ a(u,v) = \lambda b(u,v), \ \forall \ v \in H_0^1(0,1), \end{cases}$$

where

$$a(u,v) = \int_0^1 (au'\bar{v}' + cu\bar{v})dx$$

and

$$b(u,v) = \int_0^1 u\bar{v}\,dx$$

are bounded bilinear forms in $H_0^1 \times H_0^1$. An alternate formulation is

$$(3.24) \quad \begin{cases} \text{Seek } \lambda,\ 0 \neq u \in L_2(0,1) \text{ satisfying} \\ a_1(u,v) = \lambda b_1(u,v),\ \forall\, v \in H^2(0,1) \cap H_0^1(0,1), \end{cases}$$

where

$$a_1(u,v) = \int_0^1 u[-(a\bar{v}')' + c\bar{v}]dx$$

and

$$b_1(u,v) = \int_0^1 u\bar{v}\,dx$$

are bounded bilinear forms on $L_2 \times [H^2(0,1) \cap H_0^1(0,1)]$. (3.23) and (3.24) are equivalent in the sense that (λ,u) is an eigenpair of one if and only if it is an eigenpair of the other.

Another formulation is obtained as follows. If we let $\sigma = au'$, then (3.22) can be written as a first order system of equations,

$$(3.25) \quad \begin{cases} -\sigma' + cu = \lambda u \\ u' - \dfrac{\sigma}{a} = 0 \\ u(0) = u(1) = 0. \end{cases}$$

(3.25) can then be given the variational formulation,

$$(3.26) \quad \begin{cases} \text{Seek } \lambda,(\sigma,u) \in L_2(0,1) \times H_0^1(0,1) \text{ satisfying} \\ a_2(\sigma,u,\psi,v) = \lambda b_2(\sigma,u,\psi,v),\ \forall(\psi,v) \in L_2(0,1) \times H_0^1(0,1), \end{cases}$$

where

$$a_2(\sigma,u,\psi,v) = \int_0^1 (\sigma\bar{v}' + cu\bar{v} + u'\bar{\psi} - \frac{\sigma\bar{\psi}}{a})dx$$

and

$$b_2(\sigma,u,\psi,v) = \int_0^1 u\bar{v}dx.$$

a_2 and b_2 are bounded bilinear forms on $L_2 \times H_0^1$. (3.22) and (3.25) are equivalent in the sense that if (λ,u) is an eigenpair of (3.22) and $\sigma = au'$, then $(\lambda,(u,\sigma))$ is an eigenpair of (3.25), and if $(\lambda,(\sigma,u))$ is an eigenpair of (3.25), then (λ,u) is one of (3.22) and $\sigma = au'$. (3.25) and (3.26) are called <u>mixed</u> <u>formulations</u> of the eigenvalue problem (3.22); see Section 11. We can also write (3.22) in the form

$$(3.27) \quad \begin{cases} (\sigma,u) \in L_2(0,1) \times H_0^1(0,1), \quad (\sigma,u) \neq (0,0) \\[2mm] A(\sigma,\psi) + \overline{B(\psi,u)} = 0, \quad \forall\ \psi \in L_2(0,1) \\[2mm] B(\sigma,v) - \int_0^1 cu\bar{v}\ dx = \int_0^1 -\lambda\ u\bar{v}\ dx, \quad \forall\ v \in H_0^1, \end{cases}$$

where

$$A(\sigma,\psi) = \int_0^1 \frac{\sigma\bar{\psi}}{a}\ dx$$

and

$$B(\sigma,v) = -\int_0^1 \sigma\bar{v}'\ dx.$$

In Chapter III we will consider further examples of variational formulations and show how they can be used to define a variety of finite methods.

Section 4. Properties of Eigenvalue Problems

In this section we discuss the basic properties of eigenvalue problems. As in Section 3 this discussion will be in terms of 2nd order elliptic eigenvalue problems.

We thus consider the problem (3.1) in variational form:

(4.1)
$$\begin{cases} \text{Seek } \lambda, \ 0 \neq u \in H \ \text{ satisfying} \\ a(u,v) = \lambda b(u,v), \ \forall \ v \in H, \end{cases}$$

where $H = H_0^1(\Omega)$ for Dirichlet boundary conditions and $H = H^1(\Omega)$ for Neumann conditions. The forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are assumed to satisfy

(4.2)
$$|a(u,v)| \leq C_1 \|u\|_{1,\Omega} \|v\|_{1,\Omega}, \ \forall \ u,v \in H,$$

(4.3)
$$|b(u,v)| \leq C_2 \|u\|_{0,\Omega} \|v\|_{0,\Omega}, \ \forall \ u,v \in H,$$

and

(4.4)
$$\text{Re } a(u,u) \geq \alpha \|u\|_{1,\Omega}^2, \ \forall \ u \in H,$$

where $\alpha > 0$. Sufficient conditions for (4.2) − (4.4) to hold were given in Section 3; cf. (3.15) − (3.17).

For the study of (4.1) it is useful to introduce the operator $T : H^0(\Omega) \longrightarrow H$ defined by

(4.5)
$$\begin{cases} Tf \in H \\ a(Tf,v) = b(f,v), \ \forall \ v \in H. \end{cases}$$

T is the solution operator for the boundary value (source) problem

(4.6)
$$\begin{cases} Lu = df \ \text{ in } \ \Omega \\ Bu = 0 \ \text{ on } \ \Gamma, \end{cases}$$

i.e., $u = Tf$ solves (4.6). Thus T is the inverse of the dif-

ferential operator L, considered on functions that satisfy the boundary conditions. It follows immediately from (4.2) - (4.4) and the Riesz representation theorem in the special case in which $a(\cdot,\cdot)$ is an inner product on H or the Lax-Milgram theorem (Lax and Milgram [1954]) in the general case, that (4.5) has a unique solution Tf for each $f \in H^0(\Omega)$ and that

(4.7) $$\|Tf\|_{1,\Omega} \leq \frac{C_2}{\alpha}\|f\|_{0,\Omega}, \quad \forall\ f \in H^0(\Omega),$$

i.e., $T : H^0(\Omega) \longrightarrow H$ is bounded. In Section 2 we noted that H is compactly embedded in $H^0(\Omega)$, provided Γ is Lipschitz continuous (Rellich's theorem). From this fact and (4.7) we see that $T : H^0(\Omega) \longrightarrow H^0(\Omega)$ is a compact operator. We can also view T as an operator on H; we will, in fact, mainly consider T on H. Another application of Rellich's theorem shows that $T : H \longrightarrow H$ is compact.

It follows immediately from (4.1) and (4.5) that (λ,u) is an eigenpair of (4.1) (or of L) if and only if

$$Tu = \frac{1}{\lambda}u, \quad u \neq 0,$$

i.e., if and only if $(\mu = \lambda^{-1},u)$ is an eigenpair of T. Through this correspondence, properties of the eigenvalue problem (4.1) can be derived from the spectral theory for compact operators. A complete development of this theory can be found in Dunford and Schwartz [1958, 1963]. We now give a brief sketch of it; a slightly more complete treatment is given in Section 6. We present this theory under the assumption that the space H is complex. This leads to the simplest general statement of the theory. In the special case in which T is selfadjoint, H can be taken to be

43

real or complex. We will specialize to the selfadjoint case later.

Denote by $\rho(T)$ the resolvent set of T, i.e., the set

$$\rho(T) = \{z : z \in C \equiv \text{the complex numbers}, \ (z-T)^{-1}$$
$$\text{exists as a bounded operator on } H\},$$

and by $\sigma(T)$ the spectrum of T, i.e., the set $\sigma(T) = C \setminus \rho(T)$. $\sigma(T)$ is countable with no nonzero limit points; nonzero numbers in $\sigma(T)$ are eigenvalues; and if zero is in $\sigma(T)$, it may or may not be an eigenvalue. Let $0 \neq \mu \in \sigma(T)$. The space $N(\mu-T)$ of eigenvectors corresponding to μ is finite dimensional; its dimension is called the (geometric) multiplicity of μ.

Now suppose L is formally selfadjoint. Then it follows immediately from their definitions that $a(u,v)$ and $b(u,v)$ satisfy

(4.8a) $\qquad\qquad a(u,v) = a(v,u), \ \forall \ u,v \in H,$

(4.8b) $\qquad\qquad b(u,v) = b(v,u), \ \forall \ u,v \in H^0(\Omega),$

i.e., a and b are symmetric (or Hermitian) forms. Thus from (4.2) – (4.4) we see that $a(u,v)$ is an inner product on H that is equivalent to $((u,v))_{1,\Omega}$. In a similar way we see that $b(u,v)$ is an inner product on $H^0(\Omega)$ that is equivalent to $(u,v)_{0,\Omega}$ (recall that $d(x)$ is bounded above and is bounded below by a positive constant). It follows from (4.8) that

(4.9a) $\qquad\qquad a(Tu,v) = a(u,Tv), \ \forall \ u,v \in H,$

(4.9b) $\qquad\qquad b(Tu,v) = b(u,Tv), \ \forall \ u,v \in H^0(\Omega),$

i.e., T, considered as an operator on H, is selfadjoint with respect to $a(u,v)$, and, considered as an operator on $H^0(\Omega)$, is selfadjoint with respect to $b(u,v)$. (We have previously noted in

(3.10) that $b(\frac{1}{d} Lu, v) = b(u, \frac{1}{d} Lv)$ if $u = v = 0$ on Γ or if $\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = 0$ on Γ, provided L is formally selfadjoint.)

From the fact that T is selfadjoint on H it follows that the eigenvalues of T are real and the eigenfunctions corresponding to distinct eigenvalues are orthogonal with respect to $a(u,v)$. We noted above that T is compact on H and it follows from (4.5) that T is positive definite. Thus T will have a countably infinite sequence of eigenvalues

$$0 \swarrow \cdots \leq \mu_2 \leq \mu_1$$

and associated eigenfunctions

$$u_1, u_2, \ldots,$$

which satisfy

$$a(u_i, u_j) = \lambda_i b(u_i, u_j) = \delta_{ij}.$$

It is further known that the eigenfunctions are complete in $L_2(\Omega)$, i.e., that

(4.10)
$$u = \sum_{j=1}^{\infty} c_j u_j, \quad \forall u \in L_2(\Omega),$$

where

(4.11)
$$c_j = b(u, u_j) = \int_\Omega du\bar{u}_j \, dx,$$

and convergence is in either the L_2-norm or the norm $\|u\|_b = \sqrt{b(u,u)}$. (4.11) converges in the H-norm for $u \in H$.

Now the spectral properties of (4.1) (or of L) can be inferred from these facts by recalling that the eigenvalues of (4.1) (or L) are the reciprocals of those of T and that (4.1) and T have the same eigenfunctions. Thus, if L is formally

selfadjoint, then (4.1) (or L) has eigenvalues

$$(4.12) \qquad\qquad 0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow +\infty$$

and corresponding eigenfunctions

$$(4.13) \qquad\qquad u_1, u_2, \ldots$$

satisfying

$$(4.14) \qquad\qquad a(u_i, u_j) = \lambda_i b(u_i, u_j) = \delta_{ij}.$$

In the sequence $\lambda_1, \lambda_2, \ldots,$ eigenvalues are repeated according to their (geometric) multiplicity. The properties of eigenvalues and eigenfunctions used in Section 1 in the discussion of separation of variables (cf. (1.11) - (1.15)) all follow from the properties we have sketched here.

Although our discussion has been in terms of 2nd order elliptic problems, it it immediate that the results hold for any eigenvalue problem in variational form provided the bilinear forms are symmetric and satisfy (4.2) - (4.4). We will refer to this as the selfadjoint, positive definite case. In Section 8 this, as well as a more general, class of variationally formulated eigen-value problems is discussed.

Remark 4.1. The eigenvalues of selfadjoint eigenvalue problems are real and the eigenfunctions may be taken to be real. Thus these problems may be formulated in terms of real function spaces. Nonselfadjoint eigenvalue problems, on the other hand, may have complex eigenvalues and complex eigenfunctions, and are formulated in terms of complex spaces.

We end this section with a discussion of the regularity of the eigenfunctions of the 2nd order elliptic operator L. L is

46

not assumed to be formally selfadjoint here.

Theorem 4.1. Suppose for $k \geq 2$,

· $\Gamma = \partial\Omega$ is of class C^k,

· $a_{ij}, b_i \in C^{k-1}(\bar{\Omega})$, and

· $c, d \in C^{k-2}(\bar{\Omega})$.

Then all eigenfunctions of L (see (3.2)) lie in $H^k(\Omega)$ and

$$|u_j|_{k,\Omega} \leq C\lambda_j^{k/2} |u_j|_{0,\Omega}, \quad j = 1, 2, \ldots .$$

Proof. This result is a direct consequence of standard results on the regularity of solutions of elliptic boundary value problems. In particular, we refer to Agmon [1965, Theorem 9.8].

Theorem 4.2. Suppose

· Γ is of class C^v, and

· $a_{ij}, b_i, c, d \in C^v(\bar{\Omega})$.

Then $u_j \in C^v(\bar{\Omega})$ for $j = 1, 2, \ldots .$

Proof. This result follows directly from Theorem 4.1.

Theorem 4.3. Suppose

· $\Gamma = \partial\Omega$ is analytic, and

· a_{ij}, b_i, c, d are analytic on $\bar{\Omega}$.

Then u_j is analytic on $\bar{\Omega}$ for each j.

Proof. For a proof of this result see Morrey [1966, Section 5.7].

In practice most of the domains of interest have piecewise analytic boundaries. Let us mention a result for such domains.

Theorem 4.4. Suppose

- $\Omega \subset R^2$,

- Γ is piecewise analytic, and

- a_{ij}, b_i, c, d are analytic on $\bar{\Omega}$ - (vertices of $\bar{\Omega}$).

Then every eigenfunction of L is analytic in $\bar{\Omega}$ - U (vertices),

and moreover, belongs to the space $\mathcal{B}_\beta^2(\Omega)$, for properly chosen β.

Proof. This theorem follows from the results in Babuška and Guo

[1987a]. $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ ⊐

Remark 4.2. Assume that $Lu = -\Delta u$, Ω is a polygon, and the boun-

dary conditions are of Dirichlet type. If Ω is a convex poly-

gon, then the eigenfunctions $u \in H^2(\Omega)$, and if Ω is a noncon-

vex polygon, then $u \in \tilde{H}^k(\Omega) \cap H_0^1(\Omega)$, where $k = \frac{\pi}{\alpha}+1$, with α

the maximal interior angle of the vertices of Ω.

For a comprehensive treatment of regularity results for prob-

lems on domains with corners, we refer to Grisvard [1985].

Section 5.   A Brief Overview of the Finite Element Method for

Eigenvalue Approximation

In this section we give a brief overview of the use of finite element methods for approximating eigenvalues and eigenfunctions of differential operators. We will restrict the discussion to a simple model problem in one dimension and its approximation by the simplest type of finite element method.

Consider the selfadjoint eigenvalue problem

$$(5.1) \quad \begin{cases} (Lu)(x) = -(a(x)u')' + c(x)u = \lambda d(x)u, \ 0 < x < \ell \\ u(0) = u(1) = 0, \end{cases}$$

where $a \in C^1[0,\ell]$, $c,d \in C^0[0,\ell]$,   and

$$0 < a_0 \leq a(x), \ 0 \leq c(x), \quad \text{and} \quad 0 < d_0 \leq d(x) \quad \text{for} \quad 0 \leq x \leq \ell$$

(cf. (3.1) - (3.4)). As indicated in Section 3, this problem has the variational characterization

$$(5.2) \quad \begin{cases} u \in H_0^1(0,\ell) \\ a(u,v) = \lambda b(u,v), \ \forall \ v \in H_0^1(0,\ell), \end{cases}$$

where

$$a(u,v) = \int_0^\ell a(x)u'v'dx$$

and

$$b(u,v) = \int_0^\ell duvdx.$$

(5.1) (or (5.2)) has a sequence of eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow +\infty$$

49

and corresponding eigenfunctions

$$u_1, u_2, \ldots$$

satisfying

$$\lambda_i \int_0^\ell d(x) u_i u_j dx = \delta_{ij}.$$

On $[0, \ell]$ consider an arbitrary mesh

$$\Delta = \{0 = x_0 < x_1 < \ldots < x_n = \ell\},$$

where $n = n(\Delta)$ is a positive integer, and let

$$S_h = \{u : u \in C[0, \ell], \ u(0) = u(\ell) = 0,$$

$$u \text{ is linear on } I_j, \ j = 1, \ldots, n\},$$

where $h_j = x_j - x_{j-1}$ and $I_j = (x_{j-1}, x_j)$ for $j = 1, \ldots, n$ and $h = h(\Delta) = \max_j h_j$. S_h is an $(n-1)$-dimensional subspace of $H_0^1(0, \ell)$. The pairs (λ, u) have been characterized in (5.2) as eigenpairs of the bilinear form $a(u, v)$ relative to the form $b(u, v)$ over the space $H_0^1(0, \ell) \times H_0^1(0, \ell)$. We now consider eigen-pairs of $a(u, v)$ relative to $b(u, v)$ over the space $S_h \times S_h$, i.e., we consider the eigenvalue problem,

(5.3)
$$\begin{cases} \text{Seek } \lambda_h, \ 0 \neq u_h \in S_h \text{ satisfying} \\ a(u_h, v) = \lambda_h b(u_h, v), \ \forall \ u \in S_h, \end{cases}$$

and then view the eigenpairs of (5.3) as approximations to those of (5.2). (λ_h, u_h) is called a finite element (Galerkin) approxi-mation to (λ, u). A wide variety of finite element methods for eigenvalue problems will be introduced and analyzed in Chapters III. Here we will outline the general features of these methods by examining the method (5.3) as it applies to (5.1).

Since S_h is finite dimensional, (5.3) is equivalent to a generalized matrix eigenvalue problem. In fact, if $\phi_1, \ldots, \phi_{n-1}$ is a basis for S_h, then it is easily see that $\left(\lambda_h, u_h = \sum_{j=1}^{n-1} z_j \phi_j\right)$ is an eigenpair of (5.3) if and only if

$$(5.4) \qquad \begin{cases} A\vec{z} = \lambda_h B\vec{z} \\ \vec{z} \neq 0, \end{cases}$$

where $z = (z_1, \ldots, z_{n-1})^T$ and

$$A = (A_{ij}), \quad \text{with} \quad A_{ij} = a(\phi_j, \phi_i)$$

$$B = (B_{ij}), \quad \text{with} \quad B_{ij} = b(\phi_j, \phi_i).$$

(5.3) (respectively, (5.4)) has eigenvalues

$$0 < \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{n-1,h}$$

and corresponding eigenfunctions

$u_{1,h}, \ldots, u_{n-1,h}$ (respectively, $\vec{z}_{j,h} = (z_{j,1,h}, \ldots, z_{j,n-1,h})^T$,

$$j = 1, \ldots, n-1),$$

satisfying

$$\lambda_{i,h} \int_0^\ell du_{i,h} u_{j,h} \, dx = \delta_{ij} \quad (\text{respectively,} \quad \lambda_{i,h} \vec{z}_{i,h}^T B \vec{z}_{j,h} = \delta_{ij}).$$

We further note that if we choose as basis functions the usual hat functions determined by

$$\phi_i(x_j) = \delta_{ij},$$

then A and B are sparse; in fact, they are tridiagonal. We easily see that the three nonzero diagonals are given by

$$(5.5a) \quad A_{i,i+1} = -h_i^{-1}h_{i+1}^{-1}\int_{x_i}^{x_{i+1}} a(x)dx + h_i^{-1}h_{i+1}^{-1}\int_{x_i}^{x_{i+1}}(x_{i+1}-x)(x-x_i)c(x)dx$$

$$(5.5b) \quad A_{ii} = h_i^{-2}\int_{x_{i-1}}^{x_i} a(x)dx + h_{i+1}^{-2}\int_{x_i}^{x_{i+1}} a(x)dx$$

$$+ h_i^{-2}\int_{x_{i-1}}^{x_i} (x-x_{i-1})^2 c(x)dx$$

$$+ h_{i+1}^{-2}\int_{x_{i-1}}^{x_{i+1}} (x_{i+1}-x)^2 c(x)dx,$$

$$(5.5c) \quad A_{i-1,i} = -h_{i-1}^{-1}h_i^{-1}\int_{x_{i-1}}^{x_i} a(x)dx$$

$$+ h_{i-1}^{-1}h_i^{-1}\int_{x_{i-1}}^{x_i} (x-x_i)(x-x_{i-1})c(x)dx,$$

$$(5.6a) \quad B_{i,i+1} = h_i^{-1}h_{i+1}^{-1}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)(x-x_i)d(x)dx,$$

$$(5.6b) \quad B_{i,i} = h_i^{-2}\int_{x_{i-1}}^{x_i} (x-x_{i-1})^2 d(x)dx + h_{i+1}^{-2}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)^2 d(x)dx$$

$$(5.6c) \quad B_{i-1,i} = h_{i-1}^{-1}h_i^{-1}\int_{x_{i-1}}^{x_i} (x_i-x)(x-x_{i-1})d(x)dx.$$

Now we specialize (5.1) to the vibrating string problem discussed in Section 1, i.e., we let $a(x) = p =$ the tension of the string, $c(x) = 0$, and $d(x) = r =$ the density of the string. We also suppose the mesh is uniform, i.e., we let $x_j = i\ell n^{-1}$; we then have $h = h_i = \ell n^{-1}$. It is easily seen from (5.5a) - (5.6c)

52

that

$$
(5.7) \qquad A = ph^{-1} \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}
$$

and

$$
(5.8) \qquad B = \frac{rh}{6} \begin{bmatrix} 4 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & 1 & 4 & 1 \\ & & & & 1 & 4 \end{bmatrix}.
$$

If the integrals defining the B_{ij} are approximated by the trapezoid quadrature rule, then instead of the matrix B we would obtain the matrix

$$
(5.9) \qquad \tilde{B} = rhI
$$

and instead of (5.4) we would have

$$
(5.10) \qquad A\vec{z} = \tilde{\lambda}\tilde{B}\vec{z}.
$$

We finally note that the eigenvalues and eigenvectors of (5.4) and (5.10) can, in this special case, be explicitly found. The eigenvalues of (5.4) are given by

$$
(5.11) \quad \lambda_{j,h} = 6h^{-2}(1- \cos \frac{j\pi h}{\ell})(2+ \cos \frac{j\pi h}{\ell})^{-1}pr^{-1}, \quad j = 1,2,\ldots,n-1,
$$

and those of (5.10) by

$$
(5.12) \qquad \tilde{\lambda}_{j,h} = 2h^{-2}(1 - \cos \frac{j\pi h}{\ell}) pr^{-1}, \quad j = 1,2,\ldots,n-1.
$$

The unnormalized eigenvectors of both problems are given by

$$z_{j,h} = (z_{j,1,h}, \ldots, z_{j,n-1,h})^T,$$

(5.13)

where

$$z_{j,k,h} = \sin \frac{j\pi k h}{\ell}, \quad j,k = 1,2,\ldots,n-1.$$

(5.14)

The eigenvalues and eigenfunctions of (5.1), in this case, are given by

$$\lambda_j = \frac{j^2 \pi^2 p}{\ell^2 r}, \quad j = 1,2,3,\ldots$$

(5.15)

and

$$u_j(x) = \sqrt{\frac{2}{\ell r}} \sin \frac{j\pi x}{\ell}, \quad j = 1,2,\ldots .$$

(5.16)

From (5.11) and (5.15) we see that

$$\lambda_{j,h} - \lambda_j = \frac{j^4 \pi^4 p}{12 r \ell^4} h^2 + \frac{j^6 \pi^6 p}{360 r \ell^6} h^4 + \ldots = 0(h^2)$$

(5.17)

and from (5.12) and (5.15) we see that

$$\lambda_j - \tilde{\lambda}_{j,h} = \frac{j^4 \pi^4 p}{12 r \ell^4} h^2 - \frac{j^6 \pi^6 p}{360 r \ell^6} h^4 + \ldots = 0(h^2).$$

(5.18)

From (5.13), (5.14), and (5.16) we see that, neglecting the normalizing factors, the eigenvector $\vec{z}_{j,h}$ consists of the values of $u_j(x)$ at $x = x_1, x_2, \ldots, x_{n-1}$.

(5.17) shows that the eigenvalue error $\lambda_{j,h} - \lambda_j$ is $0(h^2)$. Thus the small eigenvalues of (5.3) (or of (5.4)) appro::imate the eigenvalues of (5.2), but the larger ones do not since $\lambda_{j,h} - \lambda_j$ is small only if $j^2 h$ is small. If, for example, $j \sim n^{1/2}$, then $j^2 h$ is of order one and we would not expect $\lambda_{j,h} - \lambda_j$ to be small. Thus only a small percentage of the eigenvalues of (5.4) are of interest. This observation influences the selection of numerical methods for the extraction of the eigenvalues of (5.4).

We also note that (5.17) and (5.18) show that $\tilde{\lambda}_{j,h} \le \lambda_j \le \lambda_{j,h}$ for h small. It is known that $\lambda_j \le \lambda_{j,h}$ for all h; cf. (8.42).

A Physical Interpretation of the Finite Element Eigenvalue Problem (5.10)

We consider here the vibration of a weightless elastic string loaded with several point masses. Suppose we have a weightless elastic string of length ℓ loaded with $n-1$ particles of mass m at distances ℓn^{-1}, $2\ell n^{-1}, \ldots, (n-1)\ell n^{-1}$ from one end and fixed at both ends. Gravity is assumed to be negligible and the particles are assumed to move in a plane. We shall study the small free vibrations of this system of $n-1$ degree of freedom.

Let p denote the tension in the string and let $h = \ell n^{-1}$. If $q_i(t)$ denotes the vertical displacement of the ith particle, the particles being numbered from the left (see Figure 5.1), then the equation of motion for the ith particle is easily seen to be

$$(5.19) \qquad -mq_i''(t) = -p\,\frac{q_{i-1} - 2q_i + q_{i+1}}{h}, \quad i = 1, 2, \ldots, n-1,$$

where we assume $q_0 = q_n = 0$.
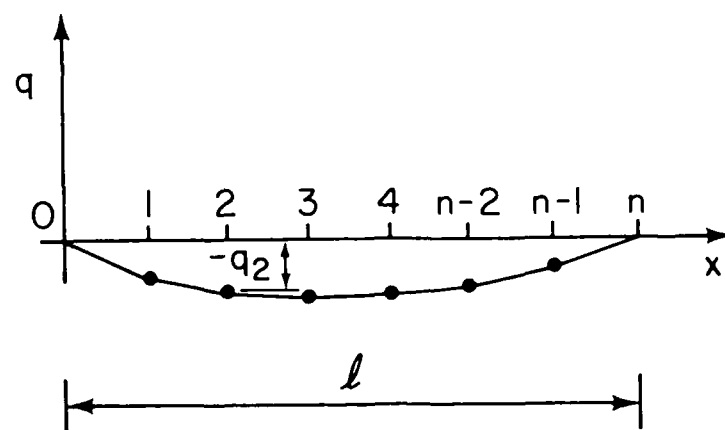
Figure 5.1. Elastic String with Point Masses.

If we seek separated solutions of the form

$$q_1(t) = z_1 q(t)$$

$$.$$
$$.$$

$$q_{n-1}(t) = z_{n-1} q(t)$$

or, in vector form,

$$\vec{q}(t) = \vec{z} q(t),$$

in which the (discrete) spatial variable j and the temporal variable t are separated, we find that

$$-m z_i q''(t) = -p \frac{z_{i-1} - 2z_i + z_{i+1}}{h} q(t)$$

or

$$-\frac{p \dfrac{z_{i-1} - 2z_i + z_{i+1}}{h}}{m z_i} = -\frac{q''(t)}{q(t)}, \quad \text{for all } i \text{ and } t.$$

Both members of this equation must equal a constant, which we denote by λ. We are thus led to seek $(\lambda, \vec{z} \neq 0)$ such that

$$\frac{p}{h}(-z_{i-1} + 2z_i - z_{i+1}) = \lambda m z_i, \quad i = 1, \ldots, n-1,$$

i.e., to seek eigenpairs $(\lambda, \vec{z})$ of the matrix

(5.20)
$$ph^{-1} \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & . & . & . & \\ & & . & . & . \\ & & & . & . & . \\ 0 & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

relative to the matrix mI, and, for each eigenvalue λ, solutions to the differential equation

(5.21) $q''(t) + \lambda q(t) = 0$, $t > 0$.

The matrix (5.20) is positive definite. Thus it has n-1
eigenvalues

$$0 < \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{n-1,h}$$

and corresponding eigenvectors $\vec{z}_{1,h}, \ldots, \vec{z}_{n-1,h}$, which satisfy

$$\lambda_{i,h} \, m\vec{z}_{i,h}^T \, \vec{z}_{j,h} = \delta_{ij}.$$

$\vec{z}_{1,h}, \ldots, \vec{z}_{n-1,h}$ thus form an orthonormal basis (i.e., are
complete) in (n-1)-dimensional space. Corresponding to $\lambda_{j,h}$,
the solutions of (5.21) are given by

$$q(t) = q_j(t) = a_j \sin \sqrt{\lambda_{j,h}}(t+\theta_j),$$

where a_j and θ_j are arbitrary. Thus the separated solutions
are given by

(5.22) $\vec{z}_{j,h} a_j \sin \sqrt{\lambda_{j,h}}(t+\theta_j)$, $j = 1, \ldots, n-1$.

As with the vibrating string, it is easily seen that all solu-
tion of (5.19) can be written as the superposition of the separated
solutions (5.22). These simple motions are called the eigenvibra-
tions. The components of the jth eigenvibration all vibrate with
some circular frequency $\sqrt{\lambda_{j,h}}$ and phase displacement $\sqrt{\lambda_{j,h}}\theta_j$,
and the components are proportional to the components of $\vec{z}_{j,h}$.
Thus $\sqrt{\lambda_{j,h}}$ is the frequency and $\vec{z}_{j,h}$ the shape of the jth
eigenvibration.

A complete discussion of the vibration of a weightless elastic
string loaded with several point masses can be found in Courant-
Hilbert [1953] and Synge and Griffith [1959].

We now draw a parallel with the finite element problem (5.10). It follows immediately from (5.7), (5.9), and (5.20) that the eigenvalue problem that we obtained, i.e., the problem of finding the eigenpairs of the matrix in (5.20) relative to mI, is identical to the eigenvalue problem (5.10) provided $m = rh = r\ell n^{-1}$. We have thus arrived at the following physical interpretation of (5.10): Consider the problem of a vibrating string with density r and tension p. Divide the total mass $r\ell$ of the string into $n-1$ particles of mass $m = r\ell n^{-1}$, which are placed at the points $x_1, \ldots, x_{n-1}$, and two particles of mass $r\ell(2n)^{-1}$, which are placed at x_0 and x_n. Then the eigenvalue problem corresponding to this system is identical to the problem (5.10) arrived at by approximating (5.2) by the finite element method (5.3), and then approximating the matrix B by $\tilde{B}$ via the trapezoid rule. Thus the finite element eigenvalue problem (5.10) is the same as the eigenvalue problem that arises when the mass of the string is "lumped" as indicated above.

The matrix A in (5.7) is called the _stiffness matrix_ and B in (5.8) is called the _mass matrix_. Because of the physical analogy we have noted, $\tilde{B}$ is called the _lumped mass matrix_ and, in contrast, B is sometimes referred to as the _consistent mass matrix_.

CHAPTER II. ABSTRACT SPECTRAL APPROXIMATION RESULTS

In this chapter we present the abstract spectral approximation results we will use in the sequel.

Section 6. Survey of Spectral Theory for Compact Operators

Since the differential operators we consider have compact inverses, our approximation results will be developed for the class of compact operators. We turn now to a survey of the spectral theory of compact operators. A complete development of this theory can be found in Dunford and Schwartz [1963, Section XI.9].

Let $A : X \longrightarrow X$ be a compact operator on a complex Banach space X with norm $\|\cdot\|_X = \|\cdot\|$. We denote by $\rho(A)$ the resolvent set of A, i.e., the set

$$\rho(A) = \{z : z \in C \equiv \text{the complex numbers}, \ (z-A)^{-1} \text{ exists}$$
$$\text{as a bounded operator on } X\},$$

and by $\sigma(A)$ the spectrum of A, i.e., the set $\sigma(A) = C \backslash \rho(A)$. For any $z \in \rho(A)$, $R_z(A) = (z-A)^{-1}$ is the resolvent operator. $\sigma(A)$ is countable with no nonzero limit points; nonzero numbers in $\sigma(A)$ are eigenvalues; and if zero is in $\sigma(A)$, it may or may not be an eigenvalue.

Let $\mu \in \sigma(A)$ be nonzero. There is a smallest integer α, called the ascent of $\mu-A$, such that $N((\mu-A)^{\alpha}) = N((\mu-A)^{\alpha+1})$, where N denotes the null space. $N((\mu-A)^{\alpha})$ is finite dimensional and $m = \dim N((\mu-A)^{\alpha})$ is called the algebraic multiplicity of μ. The vectors in $N((\mu-A)^{\alpha})$ are called generalized eigenvectors of A corresponding to μ. The order of a generalized eigenvector u is the smallest integer j such that $u \in N((\mu-A)^j)$. The

generalized eigenvectors of order 1, i.e., the vectors in

$N(\mu-A)$, are, of course, the <u>eigenvectors</u> <u>of</u> A <u>corresponding</u> <u>to</u>

μ. The <u>geometric</u> <u>multiplicity</u> <u>of</u> μ is equal to dim $N(\mu-A)$, and

is less than or equal to the algebraic multiplicity. The ascent

of $\mu-A$ is one and the two multiplicities are equal if X is a

Hilbert space and A is selfadjoint; in this case the eigenvalues

are real. If μ is an eigenvalue of A and f is a correspond-

ing eigenvector, we will often refer to (μ, f) as an <u>eigenpair</u> <u>of</u>

A.

Throughout this section we will consider a compact operator

$T : X \longrightarrow X$ and a family of compact operators $T_h : X \longrightarrow X$, $0 < h$

≤ 1, such that $T_h \longrightarrow T$ in norm as $h \searrow 0$. Let μ be a non-

zero eigenvalue of T with algebraic multiplicities m. Let Γ

be a circle in the complex plane centered at μ which lies in

$\rho(T)$ and which encloses no other points of $\sigma(T)$. The <u>spectral</u>

<u>projection</u> <u>associated</u> <u>with</u> T <u>and</u> μ is defined by

$$E = E(\mu) = \frac{1}{2\pi i} \int_\Gamma R_z(T) dz.$$

E is a projection onto the space of generalized eigenvectors

associated with μ and T, i.e., $R(E) = N((\mu-T)^\alpha)$, where R

denotes the range. For h sufficiently small, $\Gamma \subset \rho(T_h)$ and

the spectral projection

$$E_h = E_h(\mu) = \frac{1}{2\pi i} \int_\Gamma R_z(T_h) dz$$

exists, E_h converges to E in norm, and dim $R(E_h(\mu))$ =

dim $R(E(\mu))$ = m. E_h is the <u>spectral</u> <u>projection</u> <u>associated</u> <u>with</u>

T_h <u>and</u> <u>the</u> <u>eigenvalues</u> <u>of</u> T_h <u>which</u> <u>lie</u> <u>in</u> Γ and is a projec-

tion onto the direct sum of the spaces of generalized eigenvectors corresponding to these eigenvalues, i.e.,

$$R(E_h) = \sum_{\mu(h) \in \sigma(T_h), \, \mu(h) \text{ inside } \Gamma} N((\mu(h) - T_h)^{\alpha_{\mu(h)}}),$$

where $\alpha_{\mu(h)}$ is the ascent of $\mu(h) - T_h$. Thus, counting according to algebraic multiplicities, there are m eigenvalues of T_h in Γ; we denote these by $\mu_1(h), \ldots, \mu_m(h)$. Furthermore, if Γ' is another circle centered at μ with an arbitrarily small radius, we see that $\mu_1(h), \ldots, \mu_m(h)$ are all inside of Γ' for h sufficiently small, i.e., $\lim_{h \to 0} \mu_j(h) = \mu$ for $j = 1, \ldots, m$.

$R(E)$ and $R(E_h)$ are invariant subspaces for T and T_h, respectively, and $TE = ET$ and $T_h E_h = E_h T_h$. $\{R_z(T_h) : z \in \Gamma, h$ small$\}$ is bounded.

If μ is an eigenvalue of T with algebraic multiplicity m, then μ is an eigenvalue with algebraic multiplicity m of the adjoint operator T' on the dual space X'. The ascent of $\mu - T'$ will be α. E' will be the projection operator associated with T' and μ; likewise E_h' will be the projection operator associated with T_h' and $\mu_1(h), \ldots, \mu_m(h)$. If $f \in X$ and $f' \in X'$, we denote the value of the linear functional f' at f by $\langle f, f' \rangle$.

T' here is the Banach adjoint. If $X = H$ is a Hilbert space, we would naturally work with the Hilbert adjoint T^*, which acts on H. Then μ would be an eigenvalue of T if and only if $\bar{\mu}$ is an eigenvalue of T^*.

Given two closed subspaces M and N of X, we define

61

$\delta(M,N) = \sup_{\substack{x \in M \\ \|x\|=1}} \text{dist}(x,N)$ and $\hat{\delta}(M,N) = \max(\delta(M,N), \delta(N,M))$. $\hat{\delta}(M,N)$

is called the gap between M and N. The gap provides a natural way in which to formulate results on the approximation of generalized eigenvectors. We will need the following

Theorem 6.1. If dim M = dim N < ∞, then

$$\delta(N,M) \le \delta(M,N)[1-\delta(M,N)]^{-1}.$$

For a discussion of this result and the result that $\delta(N,M) = \delta(M,N)$ if X = H is a Hilbert space and $\hat{\delta}(M,N) < 1$, we refer to Kato [1958].

Section 7. Fundamental Results on Spectral Approximation

In this section we present estimates which show how the eigenvalues and generalized eigenvectors of T are approximated by those of T_h. Estimates for this type of approximation were obtained by Vainikko [1964, 1967, 1970], Bramble and Osborn [1973], and Osborn [1975]; our presentation follows Osborn [1975]. We refer also to Chatelin [1973, 1981], Grigorieff [1975 a,b,c], Chatelin and Lemordant [1978], Stummel [1977], and to the excellent and comprehensive monograph of Chatelin [1983]. Let μ be a non-zero eigenvalue of T with algebraic multiplicity m and assume the ascent of $\mu-T$ is α. Let $\mu_1(h),\ldots,\mu_m(h)$ be the eigenvalues of T_h that converge to μ.

Theorem 7.1. There is a constant C independent of h, such that

$$(7.1) \qquad \hat{\delta}(R \cdot E), R(E_h)) \le C\|(T-T_h)\|_{R(E)}\|$$

for small h, where $(T-T_h)\|_{R(E)}$ denotes the restriction of $T-T_h$ to $R(E)$.

Proof. For $f \in R(E)$ with $\|f\| = 1$ we have $\|f-E_h f\| = \|(E-E_h)f\|$ $\le \|E-E_h\|$. Thus, since E_h converges to E in norm,

$$\lim_{h \to 0} \delta(R(E),R(E_h)) = 0.$$ Using Theorem 6.1, with $M = R(E)$ and $N = R(E_h)$, we thus have

$$\delta(R(E_h),R(E)) \le \delta(R(E),R(E_h))[1-\delta(R(E),R(E_h)]^{-1}$$

$$\le 2\delta(R(E),R(E_h))$$

and hence

$$(7.2) \qquad \hat{\delta}(R(E),R(E_h)) \le 2\delta(R(E),R(E_h))$$

63

for small h.

Now for $f \in R(E)$ we have

$$\|f - E_h f\| = \|Ef - E_h f\| = \|\frac{1}{2\pi i} \int_\Gamma [R_z(T) - R_z(T_h)] f dz\|$$

$$= \|\frac{1}{2\pi i} \int_\Gamma R_z(T_h)(T - T_h) R_z(T) f dz\|$$

and hence, recalling that $R(E)$ is invariant for T and thus for $R_z(T)$,

$$(7.3) \quad \|f - E_h f\| \leq \frac{1}{2\pi} \text{ length}(\Gamma) \sup_{z \in \Gamma} \|R_z(T_h)\| \, \|(T - T_h)|_{R(E)}\| \sup_{z \in \Gamma} \|R_z(T)\| \, \|f\| .$$

As noted above, $\sup\limits_{z \in \Gamma} \|R_z(T_h)\|$ is bounded in h.   Thus from (7.2) and (7.3) we have

$$\hat{\delta}(R(E), R(E_h)) \leq C \|(T - T_h)|_{R(E)}\| ,$$

where

$$C = \frac{1}{\pi} \text{ length}(\Gamma) \sup_{\substack{z \in \Gamma \\ 0 < h}} \|R_z(T_h)\| \sup_{z \in \Gamma} \|R_z(T)\| .$$

Remark 7.1.   The proof of Theorem 7.1 also shows that

$$\|(E - E_h)|_{R(E)}\| \leq C \|(T - T_h)|_{R(E)}\| .$$

Although each of the eigenvalues $\mu_1(h), \ldots, \mu_m(h)$ is close to μ for small h,   their arithmetic mean is generally a closer approximation to μ (cf. Bramble and Osborn [1973]).   Thus we define

$$\hat{\mu}(h) = \frac{1}{m} \sum_{j=1}^{m} \mu_j(h) .$$

Out next theorem gives an estimate for $\mu - \hat{\mu}(h)$.

64

Theorem 7.2. Let $\phi_1,\ldots,\phi_m$ be any basis for $R(E)$ and let $\phi_1',\ldots,\phi_m'$ be the dual basis in $R(E')$, as defined in the proof to follow. Then there is a constant C, independent of h, such that

$$(7.4) \quad |\mu-\hat{\mu}(h)| \le \frac{1}{m}\sum_{j=1}^{m} |<(T-T_h)\phi_j,\phi_j'>| + C\,|(T-T_h)|_{R(E)}\,|(T'-T_h')|_{R(E')}\ .$$

Proof. For small h, the operator $E_h|_{R(E)} : R(E) \longrightarrow R(E_h)$ is one-to-one since $|E-E_h| \longrightarrow 0$ and $E_h f = 0$, $f \in R(E)$ implies $|f| = |Ef-E_hf| \le |E-E_h||f|$, and $E_h|_{R(E)}$ is onto since $\dim R(E_h) = \dim R(E) = m$. Thus $(E_h|_{R(E)})^{-1} : R(E_h) \longrightarrow R(E)$ is defined; we write E_h^{-1} for $(E_h|_{R(E)})^{-1}$. For h sufficiently small and $f \in R(E)$ with $|f| = 1$ we have

$$1 - |E_h f| = |Ef| - |E_h f| \le |E-E_h| \le 1/2$$

and hence $|E_h f| \ge 1/2|f|$. This implies E_h^{-1} is bounded in h. We note that $E_h E_h^{-1}$ is the identity on $R(E_h)$ and $E_h^{-1}E_h$ is the identity on $R(E)$. Now we define $\hat{T}_h = E_h^{-1}T_hE_h|_{R(E)} : R(E) \longrightarrow R(E)$.

Using the fact that $R(E_h)$ is invariant for T_h we see that $\sigma(\hat{T}_h) = \{\mu_1(h),\ldots,\mu_m(h)\}$ and that the algebraic (geometric, respectively) multiplicity of any $\mu_j(h)$ as an eigenvalue of $\hat{T}_h$ is equal to its algebraic (geometric, respectively) multiplicity as an eigenvalue of T_h. Letting $\hat{T} = T|_{R(E)}$, we likewise see that $\sigma(\hat{T}) = \{\mu\}$. Thus $\mathrm{trace}\,\hat{T} = m\mu$ and $\mathrm{trace}\,\hat{T}_h = m\hat{\mu}(h)$ and, since $\hat{T}$ and $\hat{T}_h$ act on the same space, we can write

$$(7.5) \qquad\qquad \mu-\hat{\mu}(h) = \frac{1}{m}\,\mathrm{trace}(\hat{T}-\hat{T}_h)\ .$$

Let $\varphi_1, \ldots, \varphi_m$ be a basis for $R(E)$ and let $\varphi'_1, \ldots, \varphi'_m$ be the dual basis to $\varphi_1, \ldots, \varphi_m$. Then from (7.5) we get

$$(7.6) \qquad \mu - \hat{\mu}(h) = \frac{1}{m} \, \mathrm{trace}(\hat{T} - \hat{T}_h) = \frac{1}{m} \sum_{j=1}^{m} \langle (\hat{T} - \hat{T}_h)\varphi_j, \varphi'_j \rangle.$$

Here each φ'_j is an element of $R(E)'$, the dual space of $R(E)$, but we can extend each φ'_j to all of X as follows. Since $X = R(E) \oplus N(E)$, any $f \in X$ can be written as $f = g+h$ with $g \in R(E)$ and $h \in N(E)$. Define $\langle f, \varphi'_j \rangle = \langle g, \varphi'_j \rangle$. Clearly φ'_j, so extended, is bounded, i.e., $\varphi'_j \in X'$. Now $\langle f, (\mu - T')^\alpha \varphi'_j \rangle = \langle (\mu - T)^\alpha f, \varphi'_j \rangle$ vanishes for all f. This follows from the observation that it obviously vanishes for $f \in R(E) = N((\mu - T)^\alpha)$ and it vanishes for $f \in N(E)$ since $N(E)$ is invariant for $\mu - T$. Thus we have shown that $\varphi'_1, \ldots, \varphi'_m \in R(E')$.

Using the facts that $T_h E_h = E_h T_h$ and $E_n^{-1} E_n$ is the identity on $R(E)$, we have

$$\langle (\hat{T} - \hat{T}_h)\varphi_j, \varphi'_j \rangle = \langle T\varphi_j - E_h^{-1} T_h E_h \varphi_j, \varphi'_j \rangle$$

$$(7.7) \qquad\qquad = \langle E_h^{-1} E_h (T - T_h)\varphi_j, \varphi'_j \rangle$$

$$= \langle (T - T_h)\varphi_j, \varphi'_j \rangle + \langle (E'_h E_h - I)(T - T_h)\varphi_j, \varphi'_j \rangle.$$

Let $L_h = E_h^{-1} E_h$. L_h is the projection on $R(E)$ along $N(E_h)$. Hence L'_h is the projection on $N(E_h)^\perp = R(E'_h)$ along $R(E)^\perp = N(E')$. Thus

$$(7.8) \qquad \langle (E'_h E_h - I)(T - T_h)\varphi_j, \varphi'_j \rangle = \langle (L_h - I)(T - T_h)\varphi_j, (E' - E_h)\varphi'_j \rangle.$$

From (7.8), the boundedness of L_h, and Remark 7.1 (applied to T' and $\{T'_h\}$) we have

66

$$(7.9) \quad |<(E_h'E_h-I)(T-T_h)\phi_j,\phi_j'>| \leq (\sup_h \|L_h-I\|) \|(T-T_h)\|_{R(E)}$$

$$\cdot \|(E'-E_h')\|_{R(E')} \|\phi_j\| \|\phi_j'\|$$

$$\leq C\|(T-T_h)\|_{R(E)} \|(T'-T_h')\|_{R(E')}.$$

Finally, (7.6), (7.7), and (7.9) yield the desired result.

Remark 7.2. Our treatment of the term $<(E_h'E_h-I)(T-T_h)\phi_j,\phi_j'>$, which differs from that in Osborn [1975], was suggested by Descloux, Nassif, and Rappaz [1978b].

Remark 7.3. A slight modification of the proof of Theorem 7.2 shows that for any $1 \leq \bar{i}, \bar{j} \leq m$, $|<(\hat{T}-\hat{T}_h)\phi_{\bar{j}},\phi_{\bar{i}}>|$ is bounded by $C\delta_h$, where

$$\delta_h = \sum_{i,j=1}^m |<(T-T_h)\phi_j,\phi_i'>| + \|(T-T_h)\|_{R(E)} \|(T'-T_h')\|_{R(E')}.$$

Noting that $<(\hat{T}-\hat{T}_h)\phi_j,\phi_i'>$ is a matrix representation of $\hat{T}-\hat{T}_h$, we see that

$$(7.10) \qquad \|\hat{T}-\hat{T}_h\| \leq C\delta_h.$$

Since it is immediate that

$$|\mu-\hat{\mu}(h)| = \frac{1}{m}|\text{trace}(\hat{T}-\hat{T}_h)| \leq \|\hat{T}-\hat{T}_h\|,$$

from (7.10) we get

$$(7.11) \qquad |\mu-\hat{\mu}(h)| \leq C\delta_h,$$

an estimate that is similar to, and of equal use in applications as, (7.4).

We also have

$$|\mu^{-1} - \frac{\sum_{j} \mu_j(h)^{-1}}{m}| = \frac{1}{m}|\text{trace}(\hat{T}^{-1}-\hat{T}_h^{-1})|$$

$$\leq \| \hat{T}^{-1} - \hat{T}_h^{-1} \|$$

$$= \| \hat{T}^{-1} (\hat{T} - \hat{T}_h) \hat{T}_h^{-1} \|$$

$$\leq \| \hat{T}^{-1} \| \, \| \hat{T}_h - \hat{T} \| \, \| \hat{T}_h^{-1} \|$$

$$\leq C \| \hat{T}_h - \hat{T} \| .$$

Hence we see that

(7.12)
$$\left| \mu^{-1} - \frac{\sum_j \mu_j(h)^{-1}}{m} \right| \leq C \delta_h .$$

It is also known that

(7.13)
$$| \mu - \mu_j(h) |^\alpha \leq C \| \hat{T} - \hat{T}_h \|$$

for any $1 \leq j \leq m$. Hence

(7.14)
$$| \mu - \mu_j(h) |^\alpha \leq C \delta_h .$$

(7.14) is established directly in Theorem 7.3. We note, however, that the proof of Theorem 7.3 is closely related to one of the ways of proving (7.13).

Remark 7.4. It follows immediately from (7.4) that

$$| \mu - \mu_j(h) | \leq C \| (T - T_h)|_{R(E)} \| .$$

However, the second term on the right side of (7.4) is of higher order than $\| (T - T_h)|_{R(E)} \|$, namely of order $\| (T - T_h)|_{R(E)} \| \, \| (T' - T'_h)|_{R(E)} \|$. We will also see that in a large variety of applications, $\sum_{j=1}^{m} \langle (T - T_h) \phi_j , \phi'_j \rangle$ is of higher order than $\| (T - T_h)|_{R(E)} \|$.

In addition to estimating $\mu - \mu(h)$ we may estimate $\mu - \mu_j(h)$ for each j.

Theorem 7.3. Let α be the ascent of $\mu-T$. Let $\phi_1,\ldots,\phi_m$ be any basis for $R(E)$ and let $\phi_1,\ldots,\phi_m$ be the dual basis. Then there is a constant C such that

$$(7.15) \quad |\mu-\mu_j(h)| \le C\Big\{ \sum_{i,j=1}^{m} |<(T-T_h)\phi_i,\phi_j'>| $$

$$+ |(T-T_h)|_{R(E)} |(T'-T_h')|_{R(E')} \Big\}^{1/\alpha}, \quad j = 1,\ldots,m.$$

Proof. For each h, $\mu_j(h)$ is one of the eigenvalues of $\hat{T}_h$. Let $\hat{T}_h w_h = \mu_j(h)w_h$, $|w_h| = 1$. We can choose $w_h' \in N((\mu-T')^\alpha)$ in such a way that $<w_h,w_h'> = 1$ and the norms $|w_h'|$ are bounded in h. First, using the Hahn-Banach Theorem, choose $w_h' \in R(E)'$ such that $<w_h,w_h'> = 1$ and $|w_h'| = 1$; then extend w_h' to all of X as in the proof of Theorem 7.2. w_h', so extended will be in $R(E')$ and satisfy $|w_h'| \le |E|$. Now, noting that $(T'-\mu)^\alpha w_h' = 0$, we have

(7.16)

$$|\mu-\mu_j(h)|^\alpha = |<(\mu-\mu_j(h))^\alpha w_h,w_h'>|$$

$$= |<((\mu-\mu_j(h))^\alpha-(\mu-T)^\alpha w_h,w_h'>|$$

$$= |<\sum_{j=0}^{\alpha-1} (\mu-\mu_j(h))^j (\mu-T)^{\alpha-1-j}(\mu_j(h)-T)w_h,w_h'>|$$

$$\sum_{j=0}^{\alpha-1} |\mu-\mu_j(h)|^j |<(\mu_j(h)-T)w_h,(\mu-T')^{\alpha-1-j}w_h'>|$$

$$\le \sum_{j=0}^{\alpha-1} |\mu-\mu_j(h)|^j \max_{\phi'\in R(E'),\,|\phi'|=1} |<(\mu_j(h)-T)w_h,\phi'>| \, |\mu-T'|^{\alpha-1-j} |w_h'| .$$

69

For any $\phi' \in R(E')$ with $\|\phi'\| = 1$,

(7.17)

$$|<(\mu_j(h)-T)w_h,\phi'>| = |<(\tilde{T}_h-T)w_h,\phi'>|$$

$$= |<E_h^{-1}E_h(T_h-T)w_h,\phi'>|$$

$$= |<(T-T_h)w_h,\phi'> + <(E_h^{-1}E_h-I)(T-T_h)w_h,\phi'>|$$

$$\leq |<(T_h-T)w_h,\phi'>| + C\|(T-T_h)\|_{R(E)}\|(T'-T_h')\|_{R(E')} .$$

There is obviously a constant C' such that

(7.18)
$$|<(T_h-T)w_h,\phi'>| \leq C' \sum_{i,j=1}^{m} |<(T_h-T)\phi_j,\phi_j'>|$$

for all $w_h \in R(E)$ and $\phi' \in R(E')$ with $\|w_h\| = \|\phi'\| = 1$. From (7.16) - (7.18) we get the desired result.

Theorem 7.1 shows how the generalized eigenvectors of T are approximated by those of T_h. Our next result concerns the proximity of certain elements of $R(E_h)$ to certain elements of $R(E)$. It shows, for example, that eigenvectors of T_h are close to eigenvectors of T.

Theorem 7.4. Let $\mu(h)$ be an eigenvalue of T_h such that $\lim_{h\to 0} \mu(h) = \mu$. Suppose for each h that w_h is a unit vector satisfying $(\mu(h)-T_h)^k w_h = 0$ for some positive integer $k \leq a$. Then, for any integer ℓ with $k \leq \ell \leq a$, there is a vector u_h such that $(\mu-T)^\ell u_h = 0$ and

(7.19)
$$\|u_h-w_h\| \leq C\|(T-T_h)\|_{R(E)}^{(\ell-k+1)/a} .$$

Proof. Since $N((\mu-T)^\ell)$ is finite dimensional, there is a closed subspace M of X such that $X = N((\mu-T)^\ell) \oplus M$. For y

70

$R((\mu-T)^{\ell})$, the equation $(\mu-T)^{\ell}x = y$ is uniquely solvable in M. Thus $(\mu-T)^{\ell}|_M : M \longrightarrow R((\mu-T)^{\ell})$ is one-to-one and onto. Hence $(\mu-T)^{\ell}|_M^{-1} : R((\mu-T)^{\ell}) \longrightarrow M$ exists and, by the closed graph theorem, is bounded. Thus there is a constant C such that $\|f\| \le C\|(\mu-T)^{\ell}f\|$ for all $f \in M$.

Set $u_h = Pw_h$, where P is the projection on $N((\mu-T)^{\ell})$ along M. Then $(\mu-T)^{\ell}u_h = 0$ and $w_h-u_h \in M$, and hence

$$(7.20) \qquad \|w_h-u_h\| \le C\|(\mu-T)^{\ell}(w_h-u_h)\|$$

By Theorem 7.1 there are vectors $\tilde{u}_h \in R(E)$ such that

$$\|w_h-\tilde{u}_h\| \le C'\|(T-T_h)|_{R(E)}\|.$$

Hence there is a constant C'' such that

$$
\begin{aligned}
\|[(\mu-T)^{\ell} - (\mu-T_h)^{\ell}]w_h\| & \\
(7.21) \qquad = \Big\| \sum_{j=0}^{\ell-1} (\mu-T_h)^j(T-T_h)(\mu-T)^{\ell-j-1}[(w_h-\tilde{u}_h) + \tilde{u}_h]\Big\| & \\
\le C''\|(T-T_h)|_{R(E)}\|. &
\end{aligned}
$$

Since $k \le \ell$,

$$
\begin{aligned}
(\mu-T_h)^{\ell}w_h &= \sum_{j=0}^{\ell-1} \binom{\ell}{j}(\mu-\mu(h))^j(\mu(h)-T_h)^{\ell-j}w_h \\
(7.22) \qquad &= \sum_{j=\ell-k+1}^{\ell} \binom{\ell}{j}(\mu-\mu(h))^j(\mu(h)-T_h)^{\ell-j}w_h \\
&\le C'''|\mu-\mu(h)|^{\ell-k+1}.
\end{aligned}
$$

Combining (7.20) - (7.22) we get

$$\|w_h-u_h\| \le C\|(\mu-T)^{\ell}(w_h-u_h)\|$$

$$\leq C \|(\mu-T)^{\ell} w_h\|$$

$$= C \|[(\mu-T)^{\ell} - (\mu-T_h)^{\ell}] w_h + (\mu-T_h)^{\ell} w_h\|$$

$$\leq C[C'' \|(T-T_h)\|_{R(E)}\| + C''' |\mu-\mu(h)|^{\ell-k+1}].$$

The result now follows immediately from Theorem 7.3.

Remark 7.5. If $X = H$ is a Hilbert space, we let T^* and T_h^* denote the Hilbert adjoints of T and T_h, respectively. In Theorems 7.2 and 7.3 we would let $\phi_1, \ldots, \phi_m$ be an orthonormal basis for $R(E)$ and let $\phi_j^* = E^* \phi_j$. Then $\phi_1^*, \ldots, \phi_m^* \in N((\bar{\mu}-T^*)^a)$ and trace $(\hat{T}-\hat{T}_h) = \sum_{j=1}^{m} ((\hat{T}-\hat{T}_h)\phi_j, \phi_j^*)$, where $(\cdot,\cdot) = (\cdot,\cdot)_H$ denotes the inner product on H, and with only minor modifications all the results of this section remain valid.

We end this section by specializing the results in Theorems 7.1 – 7.4 to the case where $X = H$ is a Hilbert space and T and T_h are selfadjoint. If μ is a nonzero eigenvalue of T, then, as noted above, the ascent a of $\mu-T$ is one and the algebraic and geometric multiplicities of μ are equal. Likewise the eigenvalues $\mu_j(h)$ of T_h which converge to μ have equal algebraic and geometric multiplicities. μ and $\mu_j(h)$ are, of course, real.

Thus, under the present hypotheses, Theorems 7.2 and 7.3 give the estimate

$$|\mu-\mu_j(h)| \leq C\{ \sum_{i,j=1}^{m} |((T-T_h)\phi_i, \phi_j^*)| + \|(T-T_h)|_{R(E)}\|^2 \}, \quad j = 1, \ldots, m.$$

Now consider Theorem 7.4 in the selfadjoint case. Suppose

72

$\mu(h)$ is an eigenvalue of T_h converging to μ. If w_h is a unit eigenvector of T_h corresponding to $\mu(h)$, then it follows immediately from Theorem 7.1 and the definition of $\hat{\delta}(R(E),R(E_h))$ that there is an eigenvector u_h of T corresponding to μ such that

$$u_h - w_h \leq C \, (T-T_h)|_{R(E)} \, .$$

This is Theorem 7.4 in the case $\alpha = 1$. We further note that one may assume $\|u_h\| = 1$. From Theorem 7.1 we can also conclude that if u is a unit eigenvector of T corresponding to μ then there is a unit eigenvector $w_h \in R(E_h)$ of T_h such that

$$|u - w_h| \leq C \, |(T-T_h)|_{R(E)} \, .$$

Compare the discussion of the Ritz method near the end of Section 8.

Remark 7.6. In the selfadjoint case one may assume the Hilbert space H is real (cf. Remark 4.1). Starting with a real space H we can in the usual way obtain a complex space by complexifying. Then the contour integrals $\frac{1}{2\pi i}\int_\Gamma R_z(T)dz$ and $\frac{1}{2\pi i}\int_\Gamma R_z(T_h)dz$, which are the fundamental tools in the analysis, can be introduced and the results derived. The results will be in the complex context but can immediately be translated to the real context.

Remark 7.7. Results for noncompact operators T which parallel those in this section were proved by Descloux, Nassif, and Rappaz [1978a,b]. See Subsection 13.D.

Section 8. Spectral Approximation of Variationally Formulated Eigenvalue Problems

As explained in Section 3, eigenvalue problems can be given variational formulations. For the most part, we will consider eigenvalue problems formulated in this manner. In this section we will first sketch the functional analysis framework for variationally formulated eigenvalue problems and then discuss their approximation. Results of the type presented in this section specifically Theorems 8.1 and 8.3, were proved by Babuska and Aziz [1973, Chapter 12] and Fix [1973] for the case of an eigenvalue with multiplicity one; in the general case they were proved by Kolata [1978]. Our treatment is similar to Kolata's.

Let H_1 and H_2 be complex Hilbert spaces with inner products and norms $(\cdot,\cdot)_1$ and $\|\cdot\|_1$ and $(\cdot,\cdot)_2$ and $\|\cdot\|_2$, respectively. Let $a(\cdot,\cdot)$ be a bilinear (or sesquilinear) form on $H_1 \cdot H_2$ satisfying

$$(8.1) \qquad |a(u,v)| \leq C_1 \|u\|_1 \|v\|_2, \ \forall \ u \in H_1, \ \forall \ v \in H_2,$$

$$(8.2) \qquad \inf_{\substack{u \in H_1 \\ \|u\|_1 = 1}} \ \sup_{\substack{v \in H_2 \\ \|v\|_2 = 1}} |a(u,v)| = \alpha > 0,$$

and

$$(8.3) \qquad \sup_{v \in H_1} |a(u,v)| > 0, \ \forall \ v \in H_2 \ \text{with} \ v \neq 0.$$

The Riesz representation theorem and (8.1) imply that there is a bounded linear map A from H_1 to H_2 such that $a(u,v) = (Au,v)_2$ for all $u \in H_1$, $v \in H_2$. The adjoint A' is a bounded map from H_2 to H_1 satisfying $a(u,v) = (u,A'v)_1$ for all $u \in H_1$, $v \in H_2$. (8.1), (8.2), and (8.3) imply that A is an isomor-

phism of H_1 onto H_2. In fact, in the presense of (8.1), (8.2) and (8.3) hold if and only if A is an isomorphism, cf. Babuska [1971] and Babuska and Aziz [1973, Chapter 5]. Using the fact that A is an isomorphism if and only if A' is an isomorphism we see that in the presence of (8.1), (8.2) and (8.3) hold if and only if

$$(8.4) \qquad \inf_{\substack{v \in H_2 \\ \|v\|_2 = 1}} \sup_{\substack{u \in H_1 \\ \|u\|_1 = 1}} |a(u,v)| = \sigma > 0$$

and

$$(8.5) \qquad \sup_{v \in H_2} |a(u,v)| > 0, \ \forall \ u \in H_2 \ \text{with} \ u \neq 0.$$

(8.2) and (8.3) (or (8.4) and (8.5)) are called the inf-sup conditions.

Suppose $\|\cdot\|_1'$ is a second norm on H_1 which is compact with respect to $\|\cdot\|_1$, i.e., every sequence in H_1 which is bounded in $\|\cdot\|_1$ has a subsequence which is Cauchy in $\|\cdot\|_1'$. Let $b(u,v)$ be a bilinear form on $H_1 \times H_2$ satisfying

$$(8.6) \qquad |b(u,v)| \leq C_2 \|u\|_1' \|v\|_2, \ \forall \ u \in H_1, \ v \in H_2.$$

We remark that in many applications the form $b(u,v)$ is defined on $w_1 \times w_2$, where

$$H_1 \subset w_1, \quad \text{with a compact imbedding,}$$
$$H_2 \subset w_2, \quad \text{with a bounded imbedding,}$$

and satisfies

$$(8.7) \qquad |b(u,v)| \leq C_2 \|u\|_{w_1} \|v\|_{w_2}, \ \forall \ u \in w_1, \ u \in w_2.$$

If $\|\cdot\|_1' = \|\cdot\|_{w_1}$, then it is immediate that $\|\cdot\|_1'$ is compact with respect to $\|\cdot\|_1$ and that (8.6) holds.

It is shown in Babuska [1971] and Babuska and Aziz [1973, Chapter 5] that (8.1) - (8.3) imply there are unique bounded operators $T : H_1 \longrightarrow H_1$ and $T_* : H_2 \longrightarrow H_2$ satisfying

$$(8.8) \qquad a(Tu,v) = b(u,v), \; \forall \, u \in H_1, \; \forall \, v \in H_2,$$

$$a(u,T_*v) = b(u,v), \; \forall \, u \in H_1, \; \forall \, v \in H_2.$$

Furthermore

$$(8.9) \qquad \|Tu\|_1 \leq \frac{C_2}{\sigma} \|u\|_1', \; \forall \, u \in H_1.$$

If u_j is a bounded sequence in H_1, then, since $\|\cdot\|_1'$ is compact with respect to $\|\cdot\|_1$, we know here is a subsequence u_{j_ℓ} that is Cauchy in $\|\cdot\|_1'$. It then follows immediately from (8.10), applied to $u_{j_\ell} - u_{j_k}$, that Tu_{j_ℓ} is Cauchy, and hence convergent, in H_1. Thus $T : H_1 \longrightarrow H_1$ is compact. It is immediate that $a(Tu,v) = a(u,T_*v)$. The operator T_* is related to T^*, the usual adjoint of T on H_1, by the transformation $T^* = A'T_*A'^{-1}$. T^* and T_* are compact.

A complex number λ is called an <u>eigenvalue</u> <u>of</u> <u>the</u> <u>form</u> a <u>relative</u> <u>to</u> <u>the</u> <u>form</u> b if there is a nonzero vector $u \in H_1$, called an associated <u>eigenvector</u>, satisfying

$$(8.10) \qquad a(u,v) = \lambda b(u,v), \; \forall \, v \in H_2.$$

It is easily seen from (8.8) that (λ,u) satisfies (8.10) if and only if $\lambda Tu = u$, i.e., if and only if (λ^{-1},u) is an eigenpair of the compact operator T. (8.10) is referred to as a <u>variation-ally</u> <u>posed</u> <u>eigenvalue</u> <u>problem</u> (cf. (3.18)). The notions of ascent, generalized eigenvector, and algebraic and geometric multiplicities are defined in terms of T. The generalized eigenvectors of T

76

corresponding to λ can, however, be characterized in terms of the forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. u^j is a generalized eigenvector of order $j > 1$ if and only if $a(u^j,v) = \lambda b(u^j,v) + \lambda a(u^{j-1},v)$ for all $v \in H_2$, where u^{j-1} is a generalized eigenvector of order $j-1$. Since $T_* = A'^{-1}T^*A'$, it is immediate that $\sigma(T_*) = \sigma(T^*)$ and that $N((\overline{\lambda^{-1}}-T_*)^j) = A'^{-1}(N((\overline{\lambda^{-1}}-T^*)^j))$. From this we see that the generalized eigenvectors of T_* have a similar characterization to those of T, namely, $a(u,v^j) = \lambda b(u,v^j) + \lambda a(u,v^{j-1})$ for all $u \in H_1$. In particular, $(\overline{\lambda^{-1}},v)$ is an eigenpair of T_* if and only if $a(u,v) = \lambda b(u,v)$ for all $u \in H_1$, i.e., (λ,v) is an <u>adjoint</u> <u>eigenpair</u> of (8.10).

In order to construct approximations to the eigenvalues and eigenvectors of (8.10) we select finite dimensional subspaces $S_{1,h} \subset H_1$ and $S_{2,h} \subset H_2$, indexed by a parameter h, that satisfy

$$(8.11) \qquad \inf_{\substack{u \in S_{1,h} \\ \|u\|_1 = 1}} \sup_{\substack{v \in H_{2,h} \\ \|v\|_2 = 1}} |a(u,v)| \geq \beta = \beta(h) > 0$$

and

$$(8.12) \qquad \sup_{u \in S_{1,h}} |a(u,v)| > 0, \quad \text{for each } v \in S_{2,h} \text{ with } v \neq 0.$$

We also assume

$$(8.13) \qquad \text{for every } u \in H_1, \ \lim_{h \to 0} \beta(h)^{-1} \inf_{\chi \in S_{1,h}} \|u-\chi\|_1 = 0.$$

We note that if $\dim S_{1,h} = \dim S_{2,h}$, then (8.12) follows from (8.11). We assume $\dim S_{1,h} = \dim S_{2,h}$ for the remainder of this article. $S_{1,h}$ and $S_{2,h}$ are referred to as test and trial spaces, respectively, and, if they consist of piecewise polynomial functions they are called finite element (approximation) spaces.

(8.11) is referred to as the discrete inf-sup condition.

We then consider eigenpairs of the form a relative to the form b, but now restricted to $S_{1,h} \times S_{2,h}$, i.e., pairs (λ_h, u_h), where λ_h is a number and $0 \neq u_h \in S_{1,h}$, satisfying

(8.14) $\qquad a(u_h, v) = \lambda_h b(u_h, v), \ \forall \ v \in S_{2,h}$,

and use λ_h and u_h as approximations to λ and u, respectively. (8.14) is called a __variational__ __approximation__ __method__ or __Galerkin__ __method__ for (8.10) in general and, if $S_{1,h}$ and $S_{2,h}$ consist of piecewise polynomial functions; it is called a __finite__ __element__ __method__. Since $N = \dim S_{1,h} = \dim S_{2,h} < \infty$, (8.14) is equivalent to a matrix eigenvalue problem. In fact, if $\phi_1, \ldots, \phi_N$ and $\psi_1, \ldots, \psi_N$ are bases for $S_{1,h}$ and $S_{2,h}$, respectively, then $(\lambda_h, u_h = \sum_{j=1}^{N} z_j \phi_j)$ is an eigenpair of (8.14) if and only if

(8.15) $\qquad\qquad A\vec{z} = \lambda_h B\vec{z}$,

where $z = (z_1, \ldots, z_N)^T$,

$$A = (A_{ij}), \ A_{ij} = a(\phi_j, \psi_i)$$

and

$$B = (B_{ij}), \ B_{ij} = b(\phi_j, \psi_i).$$

(λ_h, u_h) is an eigenpair of (8.14) if and only if (λ_h^{-1}, u_h) is an eigenpair of the compact operator $T_h : H_1 \longrightarrow S_{1,h}$ defined by

(8.16) $\qquad a(T_h u, v) = b(u, v), \ \forall \ u \in H_1, \ v \in S_{2,h}$.

The operator T_h can be written as $P_h T$, where P_h is the projection of H_1 onto $S_{1,h}$ defined by

(8.17) $a(P_h u, v) = a(u,v), \ \forall \ u \in H_1, \ v \in S_{2,h}.$

Using the central result in Babuška [1971] and Babuška and Aziz [1973, Chapter 6], it follows from (8.1) - (8.3), (8.11), and (8.17) that

$$\|u - P_h u\|_1 \leq (1 + \frac{C_1}{\beta(h)}) \inf_{\tau \in S_{1,h}} \|u - \tau\|_1.$$

Thus from (8.13) we see that $P_h \longrightarrow I$ pointwise. Since T is compact, $T_h = P_h T \longrightarrow T$ in norm on H_1.

Let λ be an eigenvalue of (8.10) with algebraic multiplicity m, by which we mean that λ^{-1} is an eigenvalue of T with algebraic multiplicity m. Let α = ascent of $\lambda^{-1} - T$. Since $T_h \longrightarrow T$ in norm, m eigenvalues $\lambda_1(h), \ldots, \lambda_m(h)$ of (8.14) will converge to λ. The $\lambda_j(h)$ are counted according to the algebraic multiplicities of the $\mu_j(h) = \lambda_j(h)^{-1}$ as eigenvalues of T_h. Let

(8.18) $M = M(\lambda) = \{u : u$ a generalized eigenvector of (8.10)

corresponding to $\lambda, \ \|u\|_1 = 1\}$,

(8.19) $M^* = M^*(\lambda) = \{v : v$ a generalized adjoint eigenvector of

(8.10) corresponding to $\lambda, \ \|v\|_2 = 1\}$,

and

(8.20) $M_h = M_h(\lambda) = \{u : u$ in the direct sum of the generalized

eigenspace of (8.14) corresponding to the

eigenvalues $\lambda_j(h)$ that converge to

$\lambda, \ \|u\|_1 = 1\}$,

and define

79

(8.21)
$$\varepsilon_h = \varepsilon_h(\lambda) = \sup_{u \in M(\lambda)} \inf_{\gamma \in S_{1,h}} |u - \gamma|_1$$

and

(8.22)
$$\varepsilon_h^* = \varepsilon_h^*(\lambda) = \sup_{v \in M^*(\lambda)} \inf_{\eta \in S_{2,h}} |v - \eta|_2.$$

Let $\bar{M}(\lambda) = R(E)$ and $\bar{M}_h(\lambda) = R(E_h)$.

We now state and prove four results which correspond to Theorems 7.1 - 7.4. Let α denote the ascent of $\lambda^{-1} - T$.

__Theorem 8.1.__ There is a constant C such that

(8.23)
$$\delta(\bar{M}(\lambda), \bar{M}_h(\lambda)) \leq C\beta(h)^{-1}\varepsilon_h.$$

__Theorem 8.2.__ There is a constant C such that

(8.24)
$$\left| \lambda - \left(\frac{1}{m} \sum_{j=1}^{m} \lambda_j(h)^{-1} \right)^{-1} \right| \leq C\beta(h)^{-1}\varepsilon_h\varepsilon_h^*.$$

__Theorem 8.3.__ There is a constant C such that

(8.25)
$$|\lambda - \lambda_j(h)| \leq C[\beta(h)^{-1}\varepsilon_h\varepsilon_h^*]^{1/\alpha}.$$

__Theorem 8.4.__ Let $\lambda(h)$ be an eigenvalue of (8.14) such that $\lim_{h \to 0} \lambda(h) = \lambda$. Suppose for each h that w_h is a unit vector satisfying $(\lambda(h)^{-1} - T)^k w_h = 0$ for some positive integer $k \leq \alpha$. Then, for any integer ℓ with $k \leq \ell \leq \alpha$, there is a vector u_h such that $(\lambda^{-1} - T)^\ell u_h = 0$ and

(8.26)
$$|u_h - w_h|_1 \leq C(\beta(h)^{-1}\varepsilon_h)^{(\ell - k + 1)/\alpha}.$$

__Proofs.__ The eigenvalues and generalized eigenvectors of (8.10) and (8.14) have been characterized in terms of the compact opera-

tors T and T_h and we know that $T_h \longrightarrow T$ in norm. Thus we can apply the results in Section 7, with $X = H_1$ and T and T_h as defined in (8.8) and (8.16), to estimate the eigenvalue and eigen-vector errors. Note that $M = R(E) \cap$ (unit sphere in H_1), where E is the spectral projection associated wiht T and λ^{-1}, and $M_h(\lambda) = R(E_h) \cap$ (unit sphere in H_1) where E_h is the spectral projection associated with T_h and $\lambda_j^{-1}(h)$, $j = 1,\ldots,m$. Consider first the proofs of Theorems 7.1 and 7.4. These results will follow immediately from Theorems 7.1 and 7.4, respectively, if we show that

$$(8.27) \qquad \| (T-T_h)|_{R(E)} \| \leq C\beta(h)^{-1}\varepsilon_h.$$

From Babuška [1971] and Babuška and Aziz [1973, Chapter 6] and (8.1) – (8.3), (8.6), (8.8), (8.11), and (8.16) we have

$$\| (T-T_h)u \|_1 \leq (1 + \frac{C_2}{\beta(h)}) \inf_{\iota \in S_{1,h}} \| Tu-\iota \|_1.$$

Since $M = R(E)$ is invariant for T, for $u \in R(E)$ we obtain

$$\inf_{\iota \in S_{1,h}} \| Tu-\iota \|_1 \leq \varepsilon_h \| T u \|_1.$$

(8.27) follows from these two estimates.

Now consider the proofs of Theorems 8.2 and 8.3. The right hand side of (8.4) is bounded by

$$C(\sum_{i,j=1}^{m} | ((T-T_h)\phi_j\phi_j^*)_1 | + \| (T-T_h)|_{R(E)} \| \| (T^*-T_h^*)|_{R(E^*)} \|).$$

We now show that this quantity can be bounded by $C\beta(h)^{-1}\varepsilon_h\varepsilon_h^*$. For $u \in H_1$ with $\|u\|_1 = 1$ and for $v^* \in R(E^*)$ with $\|v^*\|_1 = 1$ we have

$$((T-T_h)u,v^*)_1 = a((T-T_h)u, A'^{-1}v^*)$$

$$= a((T-T_h)u, A'^{-1}v^* - \eta)$$

$$\leq C_1 |(T-T_h)u|_1 |A'^{-1}v^* - \eta|_2, \quad \forall \eta \in S_{2,h}.$$

We have here used the definition of the operator A, (8.8), and (8.16). Recalling the A'^{-1} maps $R(E^*) = N(\overline{\lambda^{-1}-T^*})^{\alpha})$ onto $N(\overline{\lambda^{-1}-T_*})^{\alpha}) = M^*(\lambda)$, we get

$$((T-T_h)u,v^*)_1 \leq C_1 |(T-T_h)u|_1 |A'^{-1}| \varepsilon_h^*.$$

From this it is immediate that

$$|(T^*-T_h^*)v^*|_1 = \sup_{\substack{u \in H_1 \\ |u|_1 = 1}} |(u, (T^*-T_h^*)v^*)_1|$$

(8.28)

$$\leq C_1 \|T-T_h\| |A'^{-1}| \varepsilon_h^*, \quad \forall v^* \in R(E^*) \quad \text{with} \quad |v^*|_1 = 1,$$

and

$$((T-T_h)\phi_i, \phi_j^*)_1 = ((T-T_h)\phi_i, E^*\phi_j)_1$$

(8.29)

$$\leq C_1 |E^*| |(T-T_h)|_{R(E)}| |A'^{-1}| \varepsilon_h^*$$

$$\leq C \varepsilon(h)^{-1} \varepsilon_h' \varepsilon_h^*.$$

Now, using (8.27) - (8.29) we get

$$\sum_{i,j=1}^{m} |((T-T_h)\phi_j, \phi_j^*)_1| + |(T-T_h)|_{R(E)}| |(T^*-T_h^*)|_{R(E^*)}| \leq C \varepsilon(h)^{-1} \varepsilon_h' \varepsilon_h^*$$

Thus Theorem 8.2 follows from Theorem 7.2 and Theorem 8.3 from Theorem 7.3.

Remark 8.1. The proof we have given for (8.24), together with (8.12), shows that

$$(8.30) \qquad |\lambda - \hat{\lambda}(h)| \le C\beta(h)\varepsilon_h \varepsilon_h^*,$$

where

$$(8.31) \qquad \hat{\lambda}(h) = \frac{1}{m}\sum_{j=1}^{m} \lambda_j(h).$$

We end this section by specializing the results to the Ritz method for selfadjoint, positive definite problems and then presenting a lower bound for the eigenvalue error. Suppose $H_1 = H_2 = H$, $\|\cdot\|_H = \|\cdot\|$ is a real Hilbert space. Let $a(\cdot,\cdot)$ be a symmetric bilinear form on H satisfying (8.1) and

$$(8.32) \qquad a(u,u) \ge \alpha\|u\|^2, \ \forall u \in H,$$

with α a positive constant. Note that (8.1) and (8.32) imply that $a(u,u)^{1/2}$ and $\|u\|$ are equivalent norms; $a(u,u)^{1/2}$ is often called the energy norm of u. Let b be a symmetric bilinear form on W satisfying (8.7), with $W_1 = W_2 = W \supset H$ compactly, and satisfying

$$(8.33) \qquad b(u,u) > 0, \ \forall \ \text{nonzero} \ u \in H.$$

(8.32) implies (8.2) and (8.3) are satisfied. We note that (8.1) and (8.32) show that $a(\cdot,\cdot)$ is equivalent to the given inner product $(\cdot,\cdot) = (\cdot,\cdot)_H$ on H. We will now take $a(\cdot,\cdot)$ to be the inner product on H and take $\|\cdot\|_a = \sqrt{a(\cdot,\cdot)}$ to be the norm. We see that $T^* = T_* = T$. Thus T is selfadjoint; it is also positive definite. It is, of course, compact. Let $S_{1,h} = S_{2,h} = S_h \subset H$ be a family of finite dimensional spaces satisfying (8.13). In this case the variational approximation method (8.14) is called the Ritz method. (8.11), with $\beta(h) = \alpha$, and (8.12) follow from (8.32).

Under these hypotheses, the problem (8.10) has a countable sequence of eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \ldots \nearrow +\infty$$

and corresponding eigenvectors

$$u_1, \ u_2, \ u_3 \ldots,$$

which can be assumed to satisfy

(8.34) $$a(u_i, u_j) = \lambda_j b(u_i, u_j) = \lambda_j \delta_{ij}$$

(cf. Section 4). In the sequence $\{\lambda_j\}$, the λ_j are repeated according to geometric multiplicity. Furthermore, the λ_j can be characterized as various extrema of the <u>Rayleigh quotient</u>

$$R(u) = \frac{a(u,u)}{b(u,u)}.$$

We state these characterizations now.

<u>Minimum Principle</u>

(8.35) $$\lambda_1 = \min_{u \in H} R(u) = R(u_1),$$

$$\lambda_k = \min_{\substack{u \in H \\ a(u,u_i)=0,\, i=1,\ldots,k-1}} R(u) = R(u_k), \quad k = 2, 3, \ldots .$$

<u>Minimum-Maximum Principle</u>

(8.36) $$\lambda_k = \min_{\substack{V_k \subseteq H \\ \dim V_k = k}} \max_{u \in V_k} R(u) = \max_{u \in U_k = \mathrm{sp}(u_1, \ldots, u_k)} R(u), \quad k = 1, 2, \ldots .$$

<u>Maximum-Minimum Principle</u>

(8.37) $$\lambda_k = \max_{z_1, \ldots, z_{k-1} \in H} \ \min_{\substack{u \in H \\ a(u,z_i)=0,\, i=1,\ldots,k-1}} R(u)$$

$$= \min_{\substack{u \in H \\ a(u,u_i)=0, i=1,\ldots,k-1}} R(u), \qquad k = 1,2,\ldots .$$

Likewise (8.14) (with $S_{1,h} = S_{2,h} = S_h$) has a finite sequence of eigenvalues

$$0 < \lambda_{1,h} \le \lambda_{2,h} \le \ldots \le \lambda_{N,h}, \quad N = \dim S_h,$$

and corresponding eigenvectors

$$u_{1,h}, \ldots, u_{N,h},$$

which can be taken to satisfy

$$(8.38) \qquad a(u_{i,h}, u_{j,h}) = \lambda_{j,h} b(u_{i,h}, u_{j,h}) = \lambda_{j,h} \delta_{i,j}.$$

For the $\lambda_{k,h}$ we also have extremal characterizations.

Minimum Principle

$$(8.39) \qquad \lambda_{1,h} = \min_{u \in S_h} R(u) = R(u_{1,h}),$$

$$\lambda_{k,h} = \min_{\substack{u \in S_h \\ a(u,u_{i,h})=0, i=1,\ldots,k-1}} R(u) = R(u_{k,h}), \quad k = 2,\ldots,N.$$

Minimum-Maximum Principle

$$(8.40) \quad \lambda_{k,h} = \min_{\substack{V_{k,h} \subset S_h \\ \dim V_{k,h}=k}} \max_{u \in V_{k,h}} R(u) = \max_{u \in U_{k,h} = sp(u_{1,h}, \ldots, u_{k,h})} R(u),$$

$$k = 1,2,\ldots,N.$$

Maximum-Minimum Principle

$$(8.41) \qquad \lambda_{k,h} = \max_{z_{1,h},\ldots,z_{k-1,h} \in H} \min_{\substack{u \in S_h \\ a(u,z_{i,h})=0, i=1,\ldots,k-1}} R(u)$$

$$= \min_{\substack{u \in S_h \\ a(u,u_{i,h})=0, i=1,\ldots,k-1}} R(u), \qquad k = 1,\ldots,N.$$

It follows directly from the minimum and the minimum-maximum principles that

(8.42) $\qquad \lambda_k \leq \lambda_{k,h}, \quad k = 1,2,\ldots,N = \dim S_h.$

For a comprehensive treatment of such extremal characterizations of eigenvalues and their applications we refer to Courant-Hilbert [1953], Weinstein and Stenger [1972], and Weinberger [1974].

If λ_k has geometric multiplicity q, i.e., if $\lambda_k = \lambda_{k+1} = \ldots = \lambda_{k+q-1}$, then $\lambda_{k,h},\ldots,\lambda_{k+q-1,h} \searrow \lambda_k$, and combining (8.40) with Theorems 8.2 and 8.3 we see that

(8.43) $\qquad \lambda_k \leq \lambda_{j,h} \leq \lambda_k + \varepsilon_h^2(\lambda_k), \quad j = k,\ldots,k+q-1.$

(Recall that the ascent α of $\lambda_k^{-1} - T$ is one.) Regarding the approximation of eigenvectors, from Theorems 8.1 and 8.4 we see that if $w_h = u_{j,h}$, $j = k,\ldots,k+q-1$, then there is a unit eigenvector $u = u_h$ of (8.10) corresponding to λ_k such that

(8.44) $\qquad \|u-w_h\|_1 \leq \varepsilon_h,$

and if u is a unit eigenvector of (8.10) corresponding to λ_k, then there is a unit vector w_h in $sp(u_{k,h},\ldots,u_{k+q-1,h})$ such that

(8.45) $\qquad \|u-w_h\|_1 \leq C\varepsilon_h.$

If λ_k is simple, i.e., its geometric multiplicity is one, we have

(8.46) $$|u_k - u_{k,h}|_1 \leq \varepsilon_h.$$

To be more precise, if $u_1, u_2, \ldots$ satisfy (8.34), then $u_{1,h}, u_{2,h}, \ldots, u_{N,h}$ can be chosen so that (8.38) and (8.46) hold. Regarding these applications of Theorems 8.1 - 8.4, see the discussion of the selfadjoint case at the end of Section 7.

We have shown that $|\lambda - \lambda(h)| \leq C\varepsilon_h^2$, where $\lambda(h)$ is any eigenvalue that converges to λ. We now show that if λ is simple, then the error has the lower bound

$$|\lambda - \lambda(h)| \geq C\varepsilon_h^2, \quad C \text{ a positive constant.}$$

These results, together, imply that the eigenvalue error is of the same order as ε_h^2.

Theorem 8.5. (Kolata [1978]). Suppose we are in the selfadjoint, positive definite case discussed above and suppose λ is a simple eigenvalue of (8.10), i.e., it has (geometric) multiplicity one, $\lambda(h)$ is defined by the Ritz method, and $\lambda(h) \searrow \lambda$. Then there is a positive constant C such that

(8.47) $$|\lambda(h) - \lambda| \sim C\varepsilon_h^2,$$

for small h.

Proof. Let φ be a unit eigenvector of (8.10). Combining (7.6) and (7.7) yields

$$\lambda^{-1} - \lambda(h)^{-1} = a((T - T_h)\varphi, \varphi) + a((T - T_h)\varphi, L_h^*\varphi - \varphi)$$

(8.48)
$$a((T - T_h)\varphi, \varphi) - |a((T - T_h)\varphi, L_h^*\varphi - \varphi)|,$$

where L_h is the projection introduced following (7.7). Now

using (8.8), (8.16), and the definition of ε_h, we have

$$\text{(8.49)} \qquad a((T-T_h)\phi,\phi) = \lambda a((T-T_h)\phi,T\phi)$$

$$= \lambda a((T-T_h)\phi,T\phi-T_h\phi)$$

$$= \lambda \|(T-T_h)\phi\|_a^2$$

$$= \lambda^{-1}\|\phi-\lambda T_h\phi\|_a^2$$

$$\geq \lambda^{-1} \inf_{\chi \in S_h} \|\phi-\chi\|_a^2$$

$$= \lambda^{-1}\varepsilon_h^2$$

and, using (8.27) and (8.28), we have

$$|a((T-T_h)\phi,L_h^*\phi-\phi)| \leq \|(T-T_h)\phi\|_a\|L_h^*\phi-\phi\|_a$$

$$\text{(8.50)} \qquad\qquad\qquad \leq C\|(T-T_h)\phi\|_a\|(T-T_h^*)\phi\|_a$$

$$\leq C\|T-T_h\|\varepsilon_h^2.$$

Combining (8.48) - (8.50) we obtain

$$\lambda^{-1} - \lambda(h)^{-1} \geq \lambda^{-1}\varepsilon_h^2 - C\|T-T_h\|\varepsilon_h^2$$

$$= \varepsilon_h^2(\lambda^{-1}-C\|T-T_h\|)$$

$$\geq \varepsilon_h^2\lambda^{-1}/2, \quad \text{for } h \text{ small}.$$

From this we obtain

$$\lambda(h)-\lambda \geq \lambda(h)\varepsilon_h^2/2$$

$$\geq (\lambda/2)\varepsilon_h^2, \quad \text{for } h \text{ small}.$$

Remark 8.2. If λ is a multiple eigenvalue, one can prove

$$|\lambda-\lambda(h)| \leq C \inf_{\phi \in M(\lambda)} \inf_{\chi \in S_h} \|\phi-\chi\|_a^2.$$

Cf. Theorem 9.1.

Section 9.   Refined Estimates for Selfadjoint Problems

In the previous section we presented error estimates for variationally formulated problems and at the end of the section we specialized these results to the Ritz method for selfadjoint, positive definite problems. Because of the importance of this case in applications we now present an alternative analysis due to Babuška and Osborn [1987]. This analysis, which is somewhat more direct and self-contained than that in Section 8, yields stronger results than those in Section 8 in the case of multiple eigeneigenvalues.

As at the end of Section 8, we assume that $a(\cdot,\cdot)$ is a symmetric bilinear form on H satisfying (8.1) and (8.32), that $b(\cdot,\cdot)$ is a symmetric bilinear form on W satisfying (8.7), with $W_1 = W_2 = W \supset H$ compactly, and (8.33). We take $a(\cdot,\cdot)$ and $\|\cdot\|_a = \sqrt{a(\cdot,\cdot)}$ to be the inner product and norm on H and set $\|\cdot\|_b = \sqrt{b(\cdot,\cdot)}$. Then, as stated in Section 8, the eigenvalue problem (8.10), i.e., the problem

$$(9.1) \qquad \begin{cases} u \in H, \ u \neq 0 \\ a(u,v) = b(u,v), \ \forall \ v \in H \end{cases}$$

has a countable sequence of eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow +\infty$$

and corresponding eigenvectors

$$u_1, u_2, \ldots,$$

which can be chosen to satisfy

$$(9.2) \qquad a(u_i,u_j) = \lambda_j b(u_i,u_j) = \lambda_j \delta_{ij}.$$

Furthermore, any $u \in H$ can be written as

$$(9.3) \qquad u = \sum_{j=1}^{\infty} a_j u_j, \quad \text{with} \quad a_j = a(u, u_2),$$

with convergence in $\|\cdot\|_a$ (cf. (4.10) and (4.11)). We assume S_h $\subset H$ is a family of finite dimensional spaces satisfying (8.13) with $\beta(h) = a$. The eigenvalue problem (8.14) with $S_{1,h} = S_{2,h} = S_h$, i.e., the problem

$$(9.4) \qquad \begin{cases} u_h \in S_h, \ u_h \neq 0 \\ a(u_h, v) = \lambda_h b(u_h, v), \ \forall \ v \in S_h \end{cases}$$

has a finite sequence of eigenvalues

$$0 < \lambda_{1,h} \le \lambda_{2,h} \le \cdots \le \lambda_{N,h}, \quad N = \dim S_h,$$

and corresponding eigenvectors

$$u_{1,h}, \ldots, u_{N,h},$$

which can be chosen to satisfy

$$(9.5) \qquad a(u_{i,h}, u_{j,h}) = \lambda_{j,h} b(u_{i,h}, u_{j,h}) = \lambda_{i,h} \delta_{ij}.$$

The λ_k and $\lambda_{k,h}$ satisfy the extremal principles stated in Section 8.

Our analysis makes use of the following lemma that expresses a fundamental property of eigenvalue and eigenvector approximation.

Lemma 9.1. Suppose (λ, u) is an eigenpair of (9.1) with $\|u\|_b = 1$, suppose w is any vector in H with $\|w\|_b = 1$, and let $\lambda = a(w,w)$. Then

$$(9.6) \qquad \lambda - \lambda = \|w - u\|_a^2 - \lambda \|w - u\|_b^2.$$

(Note that we have assumed u and w are normalized with respect

90

to $\|\cdot\|_b$ here, whereas in (9.2) and (9.5) we assumed u_i and $u_{i,h}$ are normalized with respect to $\|\cdot\|_a$.)

Proof. By an easy calculation,

$$\|w-u\|_a^2 - \lambda\|w-u\|_b^2 = \|w\|_a^2 - 2a(w,u) + \|u\|_a^2$$
$$- \lambda\|w\|_b^2 + 2\lambda b(w,u) - \lambda\|u\|_b^2.$$

Then, since

$$\|w\|_b = \|u\|_b = 1,$$

$$\|w\|_a^2 = \tilde{\lambda},$$

$$\|u\|_a^2 = \lambda,$$

and

$$a(w,u) = \lambda b(w,u),$$

we get the desired result.

For $i = 1,2,\ldots$ suppose λ_{k_i} is an eigenvalue of (9.1) with multiplicity q_i, i.e., suppose

$$\lambda_{k_i-1} < \lambda_{k_i} = \lambda_{k_i+1} = \ldots = \lambda_{k_i+q_i-1} < \lambda_{k_i+q_i} = \lambda_{k_i+1}.$$

Here $k_1 = 1$, k_2 is the lowest index of the 2nd distinct eigen-value, k_3 is the lowest index of the 3rd distinct eigenvalue, etc. Let

$$(9.7) \qquad '_h(i,j) = \inf_{u \in M(\lambda_{k_i})} \inf_{r \in S_h} \|u-r\|_a, \quad j = 1,\ldots,q_i,$$

$$a(u,u_{k_i,h}) = \ldots = a(u,u_{k_i+j-2,h}) = 0,$$

where $M(\lambda_{k_i})$ is defined in (8.18). The restrictions $a(u,u_{k_i,h})$ $= \ldots = a(u,u_{k_i+j-2,h}) = 0$ are considered vacuous if $j = 1$.

91

Note that $\varepsilon_h(h,1) = \varepsilon_h(\lambda_i)$, where $\varepsilon_h(\lambda_i)$ is defined in (8.21).
We now estimate the eigenvalue and eigenvector errors for the
Galerkin (Ritz) method (9.4) in terms of the approximability
quantities $\varepsilon_h(i,j)$.

Theorem 9.1. There are constants C and h_0 such that

$$(9.8) \quad \lambda_{k_i+j-1,h} - \lambda_{k_i+j-1} \leq C\varepsilon_h^2(i,j), \quad \forall \ 0 < h \leq h_0, \quad j = 1,\ldots,q_i,$$
$$i = 1,2,\ldots,$$

and such that the eigenvectors $u_1, u_2, \ldots$ of (9.1) can be chosen
so that

$$(9.9) \quad \|u_{k_i+j-1,h} - u_{k_i+j-1}\|_a \leq C\varepsilon_h(i,j), \quad \forall \ 0 < h \leq h_0, \quad j = 1,\ldots,q_i,$$
$$i = 1,2,\ldots,$$

and so that (9.2) holds.

Remark 9.1. (9.8) should be compared with (8.43), (8.47), and
(8.51).

Proof. Overview of the Proof. The complete details of the proof,
which proceeds by induction, are given below. Here we provide an
overview. In Step A we give the proof for $i = 1$. The proof is
very simple in this case and rests entirely on the minimum prin-
ciple (8.39) and Lemma 9.1.

The central part of the proof is given in Step B. There we
prove the theorem for $i = 2$, proving first the eigenvalue esti-
mate (9.8) and then the eigenvector estimate (9.9). In particular,
in Steps B.1 and B.2, estimates (9.8) and (9.9), respectively, are
proved for $j = 1$. We further note that the argument used in Step
B proves the main inductive step in our proof, yielding the result
for $i = \underline{i} + 1$ on the assumption that it is true for $i \quad \underline{i}$. To

92

be somewhat more specific, the argument in Step B.1 proves (9.8) directly for any $i \geq 2$ (and $j = 1$) and that in B.2 proves (9.9) for $i = \underline{i} + 1$ (and $j = 1$) under the assumption that $\|u_{\ell,h} - u_\ell\|_a \longrightarrow 0$ as $h \longrightarrow 0$ for $\ell \leq k_{\underline{i}+1} - 1$ (cf. (9.30)).

__Details of the Proof.__ Throughout the proof we use the fact that $\varepsilon_h(i,j)$ can also be expressed as

$$\varepsilon_h(i,j) = \inf_{u \in M(\lambda_{k_i})} \quad \inf_{\tau \in S_h} \|u - \tau\|_a$$

(9.7′)
$$a(\chi, u_{k_i,h}) = \ldots = a(\chi, u_{k_i+j-2,h}) = 0$$
$$a(u, u_{k_i,h}) = \ldots = a(u, u_{k_i+j-2,h}) = 0$$

__Step A.__ Here we prove the theorem for $i = 1$.

__Step A.1.__ Suppose $\lambda_{k_1}(k_1 = 1)$ is an eigenvalue of (9.1) with multiplicity q_1, i.e., suppose $\lambda_1 = \lambda_2 = \ldots = \lambda_{q_1} < \lambda_{q_1+1}$. In this step we estimate $\lambda_{1,h} - \lambda_1$, the error between λ_1 and the approximate eigenvalue among $\lambda_{1,h}, \ldots, \lambda_{q_1,h}$ that is closest to λ_1, i.e., we prove (9.8) for $i = j = 1$. Note that

$$\varepsilon_h(1,1) = \inf_{u \in M(\lambda_1)} \inf_{\tau \in S_h} \|u - \tau\|_a$$

is the error in the approximation by elements of S_h of the most easily approximated eigenvector associated with λ_1.

From the definitions of $\varepsilon_h(1,1)$ we see that there is a $\bar{u}_h \in M(\lambda_1)$ and an $s_h \in S_h$ such that

(9.10)
$$\|\bar{u}_h - s_h\|_a = \varepsilon_h(1,1).$$

Let

$$\bar{u} = \frac{\bar{u}_h}{\sqrt{b(\bar{u}_h, \bar{u}_h)}} \quad , \quad \tilde{s}_h = \frac{s_h}{\sqrt{b(s_h, s_h)}} \quad .$$

By the minimum principle (8.39) we have

$$(9.11) \qquad \qquad \lambda_{1,h} - \lambda_1 \le a(\tilde{s}_h, \tilde{s}_h) - \lambda_1 .$$

Now apply Lemma 9.1 with $(\lambda, u) = (\lambda_1, \bar{u}_h)$, $w = \tilde{s}_h$, and $\tilde{\lambda} = a(\tilde{s}_h, \tilde{s}_h)$. We obtain

$$a(\tilde{s}_h, \tilde{s}_h) - \lambda_1 \le \|\tilde{s}_h - \bar{u}_h\|_a^2 - \lambda_1 \|\tilde{s}_h - \bar{u}_h\|_b^2$$
$$(9.12)$$
$$\|\tilde{s}_h - \bar{u}_h\|_a^2 \le C \|s_h - \bar{u}_h\|_a^2 .$$

(9.10) - (9.12) yield the desired result.

<u>Step A.2.</u> In this step we prove (9.9) for $i = j = 1$. Let $u_1, u_2, \ldots$ be eigenvectors of (9.1) satisfying (9.2). Write

$$\bar{u}_{1,h} = \sum_{j=1}^{\infty} a_j^{(1)} u_j .$$

From (9.4) and (9.11) - (9.12) we have

$$\sum \qquad \qquad \sum$$

$$\sum$$
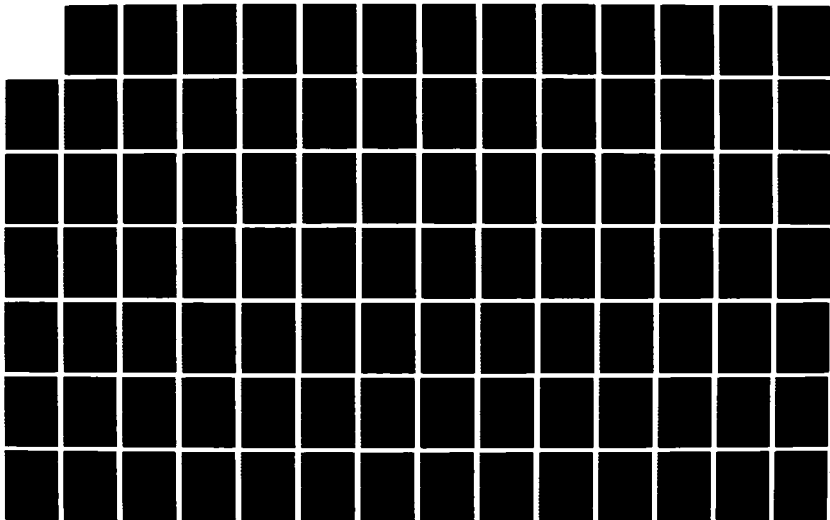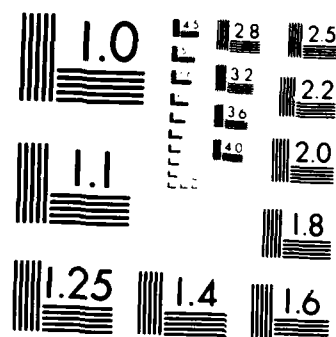
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

$$\|u_{1,h} - \sum_{j=1}^{q_1} a_j^{(1)} u_j\|_a = \left[\sum_{j=q_1+1}^{\infty} \left[a_j^{(1)} \right]^2 \right]^{1/2}$$

(9.14)

$$\leq C(1 - \lambda_1/\lambda_{q_1+1})^{-1/2} \varepsilon_h(1,1).$$

Redefining u_1 to be $\dfrac{\sum_{j=1}^{q_1} a_j^{(1)} u_j}{\|\sum_{j=1}^{q_1} a_j^{(1)} u_j\|_a}$, we easily see that $\|u_1\|_a = 1$, so that (9.2) still holds, and from (9.14) we obtain

(9.15) $\qquad\qquad \|u_{1,h} - u_1\|_a \leq C\varepsilon_h(1,1),$

as desired. Note that u_1 may depend on h.

<u>Step A.3.</u> Suppose $q_1 \geq 2$. From (9.7') we see that

(9.16) $\qquad \varepsilon_h(1,2) = \inf_{\substack{u \in M(\lambda_1) \\ a(u,u_{1,h})=0}} \inf_{\substack{\gamma \in S_h \\ a(\gamma,u_{1,h})=0}} \|u-\gamma\|_a.$

Choose $\bar{u}_h \in M(\lambda_1)$ with $a(\bar{u}_h, u_{1,h}) = 0$ and $s_h \in S_h$ with $a(s_h, u_{1,h}) = 0$ so that

(9.17) $\qquad\qquad \|\bar{u}_h - s_h\|_a = \varepsilon_h(1,2),$

and let

$$\tilde{\bar{u}}_h = \frac{\bar{u}_h}{\sqrt{b(\bar{u}_h, \bar{u}_h)}} , \quad \tilde{s}_h = \frac{s_h}{\sqrt{b(s_h, s_h)}} .$$

Since $a(s_h, u_{1,h}) = 0$, from the minimum principle (8.39), Lemma 9.1, and (9.17), we have

(9.18) $\qquad\qquad \lambda_{h,2} - \lambda_2 \leq \|\tilde{s}_h - \tilde{\bar{u}}_h\|_a^2 \leq C\varepsilon_h^2(1,2).$

This is (9.8) for $i = 1$ and $j = 2$.

<u>Step A.4</u>. In Step A.2 we redefined u_1. Now redefine $u_2, \ldots, u_{q_1}$ so that $u_1, \ldots, u_{q_1}$ are a-orthogonal. Write

$$u_{h,2} = \sum_{j=1}^{\infty} a_j^{(2)} u_j.$$

Now, proceeding as in Step A.2 and using (9.18), we have

$$(1 - \lambda_2/\lambda_{q_1+1}) \sum_{j=q_1+1}^{\infty} \left[a_j^{(2)} \right]^2 \leq \left| \sum_{j=1}^{\infty} \left[a_j^{(2)} \right]^2 (1 - \lambda_2/\lambda_j) \right|$$

$$= |a(u_{2,h}, u_{2,h}) - \lambda_2 b(u_{2,h}, u_{2,h})|$$

$$= (\lambda_{2,h} - \lambda_2) \lambda_{2,h}^{-1}$$

$$\leq C\varepsilon_h^2(1,2).$$

Thus

(9.19) $$\| u_{2,h} - \sum_{j=1}^{q_1} a_j^{(2)} u_j \|_a \leq C\varepsilon_h(1,2).$$

But by (9.15),

$$a_1^{(2)} = a(u_{2,h}, u_1)$$

$$= a(u_{2,h}, u_1 - u_{1,h})$$

(9.20) $$\leq \| u_{2,h} \|_a \| u_1 - u_{1,h} \|_a$$

$$\leq C\varepsilon_h(1,1)$$

$$\leq C\varepsilon_h(1,2).$$

Combining (9.19) and (9.20) we get

$$\| u_{2,h} - \sum_{j=2}^{q_1} a_j^{(2)} u_j \|_a \leq \| u_{2,h} - \sum_{j=1}^{q_1} a_j^{(2)} u_j \|_a + \| a_1^{(2)} u_1 \|_a$$

$$\leq C\varepsilon_h(1,2).$$

96

Redefining u_2 to be $\dfrac{\sum\limits_{j=2}^{q_1} a_j^{(2)} u_j}{\|\sum\limits_{j=2}^{q_1} a_j^{(2)} u_j\|_a}$, we see that $\|u_2\|_a = 1$ and

$a(u_1, u_2) = 0$, so that (9.2) holds and

$$(9.21) \qquad\qquad \|u_{2,h} - u_2\|_a \leq C\varepsilon_h(1,2),$$

which is (9.9) for $i = 1$, $j = 2$.

Step A.5. Continuing in the above manner we obtain the proof of (9.8) and (9.9) for $i = 1$ and $j = 1,\ldots,q_1$.

Step B. Here we prove Theorem 9.1. for $i = 2$.

Step B.1. Suppose λ_{k_2} $(k_2 = q_1+1)$ is an eigenvalue of (9.1) of multiplicity q_2. In this step we estimate $\lambda_{k_2,h} - \lambda_{k_2}$, the error between λ_{k_2} and the approximate eigenvalue among

$\lambda_{k_2,h},\ldots,\lambda_{k_2+q_2-1,h}$ that is closest to λ_{k_2}. Note that

$$(9.22) \qquad\qquad \varepsilon_h(2,1) = \inf_{u \in M(\lambda_{k_2})} \inf_{\chi \in S_h} \|u - \chi\|_a.$$

Introduce next the operators $T, T_h : H \longrightarrow H$ defined in (8.8) and (8.16), respectively, i.e., the operators defined by

$$\begin{cases} Tf \in H \\ a(Tf, v) = b(f, v), \; \forall \; v \in H \end{cases}$$

and

$$\begin{cases} T_h f \in S_h \\ a(T_h f, v) = b(f, v), \; \forall \; v \in S_h \end{cases}$$

It follows from (8.1), (8.7), (8.32), and (8.33) that T and T_h are defined and compact on H. Furthermore

$$(9.23) \qquad \|(T-T_h)f\|_a \leq C \inf_{\chi \in S_h} \|Tf - \chi\|_a.$$

We now suppose the space H and the bilinear forms a and b have been complexified in the usual manner. Let Γ be a circle in the complex plane centered at $\mu_{k_2} = \lambda_{k_2}^{-1}$, enclosing no other eigenvalues of T. Then for h sufficiently small, Γ $\Gamma \subset \rho(T_h)$ and $\text{Int } (\Gamma) \cap \sigma(T_h) = \{u_{k_2,h}, \ldots, u_{k_2+q_2-1,h}\}$, where $\lambda_{k_2+i} = \lambda_{k_2+i}^{-1}$. Also, as we have seen in Section 6, $E(\mu_{k_2})$, the spectral projections associated with T and μ_{k_2}, and $E_h(\mu_{k_2})$, the spectral projection associated with T_h and $\mu_{k_2+i,h}$, $i = 0,\ldots,q_2-1$, respectively, can be written as

$$(9.24) \qquad E(\mu_{k_2}) = \frac{1}{2\pi i} \int_\Gamma (z-T)^{-1} dz$$

and

$$(9.25) \qquad E_h(\mu_{k_2}) = \frac{1}{2\pi i} \int_\Gamma (z-T_h)^{-1} dz$$

Let $u \in R(E(\mu_{k_2}))$. Then $v_h = E_h(\mu_{k_2}) u \in R(E(\mu_{k_2}))$, and from the formulas (9.24) and (9.25) we obtain

$$\|u - v_h\|_a = \|(E(\mu_{k_2}) - E_h(\mu_{k_2}))u\|_a$$

$$= \left\| \frac{1}{2\pi i} \int_\Gamma (z-T_h)^{-1}(T-T_h)(z-T)^{-1}u \; dz \right\|$$

$$(9.26) \qquad = \left\| \frac{1}{2\pi i} \int_\Gamma (z-T_h)^{-1}(T-T_h)\frac{u}{z-\mu_{k_2}} \; dz \right\|$$

$$\leq \frac{1}{2\pi} [2\pi \; \text{rad}(\Gamma)] \sup_{z\in\Gamma} \|(z-T_h)^{-1}\| \frac{1}{\text{rad}(\Gamma)} \|(T-T_h)u\|_a$$

98

$$= (\mu_{k_2+q_2-1,h} - \mu_{k_2}+\text{rad}(\Gamma))^{-1} \|(T-T_h)u\|_a$$

$$\leq C\|(T-T_h)u\|_a .$$

(9.23) and (9.26) yield

$$\|u-v_h\|_a \leq C \inf_{\substack{\iota \in S_h}} \|Tu-\iota\|_a$$

(9.27)
$$= C \inf_{\substack{\chi \in S_h}} \|\mu_{k_2}u-\iota\|_a$$

$$\leq C \inf_{\substack{\chi \in S_h}} \|u-\chi\|_a .$$

This is an eigenvector estimate; it shows that starting from any

$u \in R(E(\mu_{k_2}))$ we can construct a $v_h = v_h(u) \in R(E_h(\mu_{k_2}))$ that

is close to u. We now use (9.27) to prove the desired eigenvalue

estimate.

By the minimum principle (8.39) we have

(9.28)
$$\lambda_{k_2,h} - \lambda_{k_2} = \inf_{\substack{v \in S_h \\ \|v\|_b=1 \\ a(v,u_{i,h})=0, \\ i=1,\ldots,k_2-1}} a(v,v) - \lambda_{k_2} .$$

Since $v_h = R(E_h(\mu_{k_2}))$, we know that $a(v_h,u_{i,h}) = 0$, i =

$1,\ldots,k_2-1$. Thus, from (9.28) we find

$$\lambda_{k_2,h} - \lambda_{k_2} \leq a\left(\frac{v_h}{\|v_h\|_b}, \frac{v_h}{\|v_h\|_b}\right) - \lambda_{k_2} .$$

Combining this with Lemma 9.1 and (9.27) we obtain

$$\lambda_{k_2,h} - \lambda_{k_2} \leq \left\|\frac{v_h}{\|v_h\|_b} - \frac{u}{\|u\|_b}\right\|_a^2 - \lambda_{k_2}\left\|\frac{v_h}{\|v_h\|_b} - \frac{u}{\|u\|_b}\right\|_b$$

$$\leq C\|v_h - u\|_a^2$$

$$\le C \inf_{\chi \in S_h} \|u - \chi\|_a^2$$

for $u \in R(E(\mu_{k_2}))$ with $\|u\|_a = 1$. Hence, using (9.22),

$$(9.29) \qquad \lambda_{k_2,h} - \lambda_{k_2} \le C \inf_{u \in M(\lambda_{k_2})} \inf_{\chi \in S_h} \|u - \chi\|_a^2$$

$$= C\varepsilon_h^2(2,1),$$

which is (9.8) for $i = 2$, $j = 1$.

__Comment on Inequality (9.29)__. A careful examination of the proof of (9.29) shows that C depends only on $\mu_{k_1} - \mu_{k_2}$ and $\mu_{k_2} - \mu_{k_3}$, but is independent of h, and that (9.29) is valid for $h \le h_0$, where h_0 is such that $h \le h_0$ implies $\Gamma \subset \rho(T_h)$, Int $(\Gamma) \cap \sigma(T_h) = \{u_{k_2,h}, \ldots, u_{k_2+q_2-1,h}\}$, and $u_{k_2} - u_{k_2+q_2-1}$ is small, say $u_{k_2} - u_{k_2+q_2-1,h} \le$ rad $(\Gamma)/2$. Note that if we were considering a family of problems depending on a parameter τ, we could bound $C = C(\tau)$ above, independent of τ, provided $\mu_{k_1}(\tau) - \mu_{k_2}(\tau)$ and $\mu_{k_2}(\tau) - \mu_{k_3}(\tau)$ were bounded away from 0, and we could bound $h_0(\tau)$ away from 0 if $\Gamma(\tau) \subset \rho(T_h(\tau))$, Int $(\Gamma(\tau)) \cap \sigma(T_h(\tau)) = \{u_{k_2,h}, \ldots, u_{k_2+q_2-1,h}\}$, and $\mu_{k_2}(\tau) - u_{k_2+q_2-1,h}(\tau) \le$ rad $(\Gamma(\tau))/2$, uniformly in τ.

__Step B.2__. Suppose, as in Step B.1, that λ_{k_2} has multiplicity q_2. We have shown in Step A.5 that we can choose the eigenvectors $u_1, u_2, \ldots$ of (9.1) so that (9.2) holds and so that

$$(9.30) \qquad \|u_{h,j} - u_j\|_a \le C\varepsilon_h(1,j), \quad j = 1, \ldots, q_1 = k_2 - 1.$$

Write

$$(9.31) \qquad u_{k_2,h} = \sum_{j=1}^{\infty} a_j^{(k_2)} u_j.$$

From (9.31) we have

$$\left| \sum_{j=1}^{\infty} \left[a^{(k_2)} \right]^2 (1 - \lambda_{k_2}/\lambda_j) \right| = \left| a(u_{k_2,h}, u_{k_2,h}) - \lambda_{k_2} b(u_{k_2,h}, u_{k_2,h}) \right|$$

$$= (\lambda_{k_2,h} - \lambda_{k_2}) \lambda_{k_2,h}^{-1},$$

which, together with (9.29), yields

$$(9.32) \quad \left| \sum_{j=1}^{k_2-1} \left[a^{(k_2)} \right]^2 (1 - \lambda_{k_2}/\lambda_j) + \sum_{j=k_2+q_2}^{\infty} \left[a^{(k_2)} \right]^2 (1 - \lambda_{k_2}/\lambda_j) \right|$$

$$\leq C \varepsilon_h^2(2,1).$$

Note that the first term inside the absolute value is negative and the second is positive. In addition

$$C_1 \leq |1 - \lambda_{k_2}/\lambda_j| \leq C_2, \quad \forall j = k_2, \; k_2+1, \; \ldots, \; k_2+q_2-1,$$

with C_1, C_2 positive numbers. Hence from (9.32) we obtain

$$(9.33) \qquad \sum_{j=1}^{k_2-1} \left[a^{(k_2)} \right]^2 \leq D_1 \varepsilon_h^2(2,1) + D_2 \sum_{j=k_2+q_2}^{\infty} \left[a^{(k_2)} \right]^2$$

and

$$(9.34) \qquad \sum_{j=k_2+q_2}^{\infty} \left[a^{(k_2)} \right]^2 \leq D_3 \varepsilon_h^2(2,1) + D_4 \sum_{j=1}^{k_2-1} \left[a^{(k_2)} \right]^2.$$

Write

$$(9.35) \qquad u_{i,h} - u_i = \sum_{j=1}^{\infty} b_{i,j} u_j, \quad i = 1, \ldots, k_2 - 1 = q_1.$$

Then, by (9.30),

$$(9.36) \quad \sum_{j=1}^{\infty} b_{i,j}^2 = \| u_{i,h} - u_i \|_a^2 \leq C \varepsilon_h^2(1,i), \quad i = 1, \ldots, k_2 - 1.$$

Next we wish to find constants $\alpha_1, \ldots, \alpha_{k_2-1}$ so that

$$(9.37) \qquad a(u_i, \sum_{j=1}^{k_2-1} \alpha_j u_{j,h}) = a_i^{(k_2)}, \quad i = 1, \ldots, k_2 - 1.$$

Using (9.35), these equations can be written as

$$a\left[u_i, \sum_{j=1}^{k_2-1} (\alpha_j u_j + \alpha_j \sum_{\ell=1}^{\infty} b_{j,\ell}, u_\ell) \right] = \alpha_i + \sum_{j=1}^{k_2-1} b_{j,i} \alpha_j$$

(9.38)
$$= a_i^{(k_2)}, \quad i = 1, \ldots, k_2 - 1.$$

Since (8.13) implies $\varepsilon_h^2(2,1) \longrightarrow 0$ as $h \longrightarrow 0$, from (9.36) we see that the $b_{j,i}$ are small for $h \leq h_0$, with h_0 sufficiently small, and hence the system (9.38) is uniquely solvable, and, moreover, there is a constant L, depending only on k_2, such that

$$(9.39) \qquad \left[\sum_{j=1}^{k_2-1} \alpha_i^2 \right]^{1/2} \leq L \left[\sum_{j=1}^{k_2-1} \left[a_i^{(k_2)} \right]^2 \right]^{1/2}.$$

Now, from (9.30) we obtain

$$|a_j^{(k_2)}| = |a(u_{k_2,h}, u_j)|$$

$$= |a(u_{k_2,h}, u_j - u_{j,h})|$$

$$\leq \|u_{k_2,h}\|_{B_0} \|u_j - u_{j,h}\|_a$$

$$= \|u_j - u_{j,h}\|_a$$

$$\leq C\varepsilon_{1,j}(h), \quad j = 1,\ldots,k_2-1.$$

Letting

$$(9.40) \qquad \rho_{k_2}^2(h) = \sum_{j=1}^{k_2-1} \varepsilon_h^2(1,j),$$

we see that

$$(9.41) \qquad \left[\sum_{j=1}^{k_2-1} \left[a_j^{(k_2)}\right]^2\right]^{1/2} \leq C\rho_{k_2}(h),$$

and thus, from (9.39)

$$(9.42) \qquad \left[\sum_{j=1}^{k_2-1} \alpha_i^2\right]^{1/2} \leq LC\rho_{k_2}(h)$$

$$\leq C\rho_{k_2}(h).$$

Now let

$$(9.43) \qquad \psi = u_{k_2,h} - \sum_{j=1}^{k_2-1} \alpha_j u_{j,h}.$$

Then $\psi \in S_h$. Furthermore, from (9.35) and (9.37) we get

$$(9.44) \qquad a(u_i,\psi) = \begin{cases} 0, & i \leq k_2 - 1 \\[2mm] a_i^{(k_2)} - \sum_{j=1}^{k_2-1} \alpha_j b_{j,i}, & i \geq k_2. \end{cases}$$

From (9.42) and (9.43),

$$\left|\, \|\psi\|_a - 1 \,\right| = \left|\, \|\psi\|_a - \|u_{k_2,h}\|_a \,\right|$$

$$\leq \|\psi - u_{k_2,h}\|_a$$

(9.45)

$$\leq \left[\sum_{j=1}^{k_2-1} \sigma_j^2\right]^{1/2}$$

$$\leq C\rho_{k_2}(h).$$

Using (9.29), (9.44), and (9.45), and the fact that $\rho_{k_2}(h)\longrightarrow 0$ as $h\longrightarrow 0$, we get

$$C\varepsilon_h^2(2,1) \geq \frac{\lambda_{k_2,h} - \lambda_{k_2}}{\lambda_{k_2,h}}$$

(9.46)

$$\geq a(u_{k_2,h}, \frac{\psi}{\|\psi\|_a}) - \lambda_{k_2} b(u_{k_2,h}, \frac{\psi}{\|\psi\|_a})$$

$$= C'\left[\sum_{\ell=k_2+q_2}^{\infty} a_\ell^{(k_2)}\left(a_\ell^{(k_2)} - \sum_{i=1}^{k_2-1}\alpha_i b_{i,\ell}\right)\left(1 - \frac{\lambda_{k_2}}{\lambda_\ell}\right)\right],$$

where $C' > 0$ and is independent of h. Combining (9.36), (9.39), (9.40), and (9.46) we obtain

$$\sum_{\ell=k_2+q_2}^{\infty}\left[a_\ell^{(k_2)}\right]^2 \leq C\Bigg[\varepsilon_h^2(2,1)$$

$$+ \sum_{\ell=k_2+q_2}^{\infty}|a_\ell^{(k_2)}|\sum_{i=1}^{k_2-1}|\alpha_i||b_{i,\ell}|\Bigg]$$

$$\leq C\left[\varepsilon_h^2(2,1) + \sum_{i=1}^{k_2-1}|\alpha_i|\sum_{\ell=k_2+q_2}^{\infty}|a_\ell^{(k_2)}||b_{i,\ell}|\right]$$

104

$$\leq C\left[\varepsilon_h^2(2,1) + \sum_{i=1}^{k_2-1} |\alpha_i| \left(\sum_{\ell=k_2+q_2}^{\infty} |a_\ell^{(k_2)}|^2\right)^{1/2},\right.$$

$$\left.\left(\sum_{\ell=k_2+q_2}^{\infty} |b_{i,\ell}|^2\right)^{1/2}\right]$$

(9.47)

$$\leq C\left[\varepsilon_h^2(2,1) + \sum_{i=1}^{k_2-1} |\alpha_i| \left(\sum_{\ell=k_2+q_2}^{\infty} |a_\ell^{(k_2)}|^2\right)^{1/2}\right.$$

$$\left.\max_{i=1,\dots k_2-1} \varepsilon_h(1,i)\right]$$

$$\leq C\left[\varepsilon_h^2(2,1) + \varepsilon_h(1,k_2-1)\sqrt{k_2-1}\left(\sum_{i=1}^{k_2-1} |\alpha_i|^2\right)^{1/2}\right.$$

$$\left.\left(\sum_{\ell=k_2+q_2}^{\infty} |a_\ell^{(k_2)}|^2\right)^{1/2}\right]$$

$$\leq C\left[\varepsilon_h^2(2,1) + \varepsilon_h(1,k_2-1)\sqrt{k_2-1}\ L\left(\sum_{i=1}^{k_2-1} \left[a_i^{(k_2)}\right]^2\right)^{1/2}\right.$$

$$\left.\left(\sum_{\ell=k_2+q_2}^{\infty} |a_\ell^{(k_2)}|^2\right)^{1/2}\right]$$

$$\leq C\left[\varepsilon_h^2(2,1) + \varepsilon_h(1,k_2-1)\left(\sum_{i=1}^{k_2-1} \left[a_i^{(k_2)}\right]^2\right)^{1/2}.\right.$$

$$\left.\left(\sum_{\ell=k_2+q_2}^{\infty} |a_\ell^{(k_2)}|^2\right)^{1/2}\right].$$

(9.47) is a quadratic inequality in $\left[\sum_{\ell=k_2+q_2}^{\infty}\left[a_\ell^{(k_2)}\right]^2\right]^{1/2}$, whose

solution yields

$$(9.48) \qquad \sum_{\ell=k_2+q_2}^{\infty}\left[a_\ell^{(k_2)}\right]^2 \leq C\varepsilon_h(1,k_2-1)\sum_{i=1}^{k_2-1}\left[a_i^{(k_2)}\right]^2 + C\varepsilon_h^2(2,1).$$

Combining (9.33) and (9.48) we get

$$\sum_{i=1}^{k_2-1}\left[a_i^{(k_2)}\right]^2 \leq D_1\varepsilon_h^2(2,1) + D_2C\varepsilon_h(1,k_2-1)\sum_{i=1}^{k_2-1}\left[a_i^{(k_2)}\right]^2 + D_2C\varepsilon_h^2(2,1),$$

and thus, since $\varepsilon_h(1,k_2-1)$ is small for h small,

$$(9.49) \qquad \sum_{i=1}^{k_2-1}\left[a_i^{(k_2)}\right]^2 \leq D_5\varepsilon_h^2(2,1).$$

Next, combining (9.34) and (9.49), we get

$$(9.50) \qquad \sum_{\ell=k_2+q_2}^{\infty}\left[a_\ell^{(k_2)}\right]^2 + D_6\varepsilon_h^2(2,1).$$

Finally, from (9.31), (9.49), and (9.50) we have

$$u_{k_2,h} - \sum_{j=k_2}^{k_2+q_2-1}a_j^{(k_2)}u_j a = \left[\sum_{j=1}^{k_2-1}\left[a_j^{(k_2)}\right]^2 + \sum_{j=k_2+q_2}^{\infty}\left[a_j^{(k_2)}\right]^2\right]^{1/2}$$

$$\leq C\varepsilon_h(2,1).$$

Redefining u_{k_2} to be $\dfrac{\sum\limits_{j=k_2}^{k_2+q_2-1}a_j^{(k_2)}u_j}{\sum\limits_{j=k_2}^{k_2+q_2-1}a_j^{(k_2)}u_j a}$, we see that $u_{k_2} a = 1$.

so that (9.2) holds, and

$$(9.51) \qquad \|u_{k_2,h} - u_{k_2}\|_a \le C\varepsilon_h(2,1).$$

This is (9.9) for $i = 2$, $j = 1$.

Comment on Estimate (9.51). In the proof of (9.51) we used (9.30), which was proved in Step A. A careful examination of the proof of (9.51) shows that we that we did not use the full strength of (9.30), but only the weaker fact that $\|u_{j,h} - u_j\|_a \longrightarrow 0$ as $h \longrightarrow 0$ for $j \le k_2 - 1$. (Cf. the Overview of the Proof.)

Step B.3. Suppose $q_2 \ge 2$. In Step B.1 we estimated $\lambda_{k_2,h} - \lambda_{k_2}$. In this step we estimate $\lambda_{k_2+1,h} - \lambda_{k_2+1}$.

We proceed by modifying problems (9.1) and (9.4) by restricting them to the spaces

$$H^{k_2,h} = \{u \in H : a(u, u_{k_2,h}) = 0\}$$

and

$$S_h^{k_2,h} = \{u \in S_h : a(u, u_{k_2,h}) = 0\},$$

respectively, i.e., we consider the problems $(9.1^{k_2,h})$ and $(9.4^{k_2,h})$ obtained by replacing H and S_h by $H^{k_2,h}$ and $S_h^{k_2,h}$ in (9.1) and (9.4), respectively. $(9.4^{k_2,h})$ has the same eigenpairs $(\lambda_{j,h}, u_{j,h})$ as does (9.4), except that the pair $(\lambda_{k_2,h}, u_{k_2,h})$ has been eliminated. $(9.1^{k_2,h})$ has eigenpairs $(\lambda_j^{k_2,h}, u_j^{k_2,h})$, which in general depend on h. Nevertheless,

$$(9.52) \qquad \lambda_{k_2+\ell}^{k_2,h} = \lambda_{k_2+\ell}, \quad \ell = 0, \ldots, q_2 - 2,$$

i.e., λ_{k_2}, the eigenvalue under consideration, is an eigenvalue of multiplicity q_2-1 for problem $(9.1^{k_2,h})$. Its eigenspace is

$$\bar{M}^{k_2,h}(\lambda_{k_2}) = \{u \in M(\lambda_{k_2}) : a(u,u_{h,k_2}) = 0\}.$$

We can now apply the argument used in Step B.1 to problems $(9.1^{k_2,h})$ and $(9.4^{k_2,h})$ and, using $(9.7')$, we obtain (cf. (9.29))

$$(9.53) \qquad \lambda_{k_2+1,h} - \lambda_{k_2+1} \leq C\varepsilon_{2,2}^2(h), \quad \text{for} \quad h < h_0.$$

Since $u_{k_2,h}$ depends on h, the problems $(9.1^{k_2,h})$ and $(9.4^{k_2,h})$ depend on h. It follows from the Comment on Inequality (9.29) with $\tau = h$ that we can apply the argument in Step B.1 obtaining C and h_0 that are independent of h. To see this, note that $\mu_{k_2}(\tau) = \mu_{k_2}$, by (9.52), $\mu_{k_3}(\tau) \leq \mu_{k_3}$, by the minimum principle, and $\mu_{k_1}(\tau) \longrightarrow \mu_{k_1}$, since $\mu_{k_1}-\mu_{k_1}(\tau) \leq \mu_{k_1}-\mu_{k_1,h}$, by the minimum-maximum principle, and $u_{k_1,h} \longrightarrow u_{k_1}$ (cf. (9.51)), and hence that $\mu_{k_1}(\tau) - \mu_{k_2}(\tau)$ and $\mu_{k_2}(\tau) - \mu_{k_3}(\tau)$ are bounded away from 0. Then note that $\Gamma(\tau) = \Gamma \subset \rho(T_h(\tau)) = \rho(T_h)-\{u_{k_2,h}\}$, Int $(\Gamma(\tau)) \cap \sigma(T_h(\tau)) = $ Int $(\Gamma) \cap (\sigma(T_h)-\{u_{k_2,h}\}) = \{u_{k_2+1,h},\cdots,u_{k_2+q_2-1,h}\}$, and $u_{k_2}(\tau)-u_{k_2+q_2-1,h}(\tau) = u_{k_2}-u_{k_2+q_2-1,h} \leq $ rad $(\Gamma(\tau))/2 = $ rad $(\Gamma)/2$.

Step B.4. Suppose $q_2 \geq 2$ as in Step B.3. Here we show that u_{k_2+1} can be chosen so that $u_{k_2+1,h} - u_{k_2+1} \, a \leq C\varepsilon_h(2,2)$. We know that

108

$$(9.54) \qquad \|u_{j,h} - u_j\|_a \leq \begin{cases} C\varepsilon_h(1,j), & j = 1,\ldots,q_1 \\ C\varepsilon_h(2,1), & j = q_1+1 = k_2 \end{cases},$$

(cf. (9.15), (9.21), and (9.51)). Assume that $u_{k_2+1},\ldots,u_{k_2+q_2-1}$ have been redefined so that (9.2) holds. Write

$$u_{k_2+1,h} = \sum_{j=1}^{\mathcal{L}} a_j^{(k_2+1)} u_j.$$

If we apply the argument used in Step B.2 to $u_{k_2+1,h}$, i.e., if we let k_2 be replaced by k_2+1 and use (9.53) instead of (9.29), we obtain

$$\left\| u_{k_2+1,h} - \sum_{j=k_2}^{k_2+q_2-1} a_j^{(k_2)} u_j \right\|_a \leq C\varepsilon_h(2,2).$$

But, by (9.54),

$$\begin{aligned}
|a_{k_2}^{(k_2+1)}| &= |a(u_{k_2+1,h}, u_{k_2})| \\
&= |a(u_{k_2+1,h}, u_{k_2} - u_{k_2,h})| \\
&\leq \|u_{k_2} - u_{k_2,h}\|_a \\
&\leq C\varepsilon_h(2,1) \\
&\leq C\varepsilon_h(2,2)
\end{aligned}$$

and hence

$$\left\| u_{k_2+1,h} - \sum_{j=k_2+1}^{k_2+q_2-1} a_j^{(k_2)} u_j \right\|_a \leq C\varepsilon_{2,2}(h).$$

Redefining u_{k_2+1} to be $\dfrac{\sum\limits_{j=k_2+1}^{k_2+q_2+1} a_j^{(k_2)} u_j}{\left\|\sum\limits_{j=k_2+1}^{k_2+q_2-1} a_j^{(k_2)} u_j\right\|_a}$, we see that

$\|u_{k_2+1}\|_a = 1$, $a(u_{k_2+1}, u_j) = 0$, $j = 1, \ldots, k_2$, so that (9.2)

holds, and

$$\|u_{k_2+1,h} - u_{k_2+1}\|_a \leq C\varepsilon_{2,2}(h),$$

which is (9.9) for $i = j = 2$.

<u>Step B.5</u>. Continuing in this manner we prove (9.8) and (9.9)

for $i = 2$ and $j = 1, \ldots, q_2$.

<u>Step C</u>. Repeating the argument in B we get (9.8) and (9.9)

for $i = 3, 4, \ldots$. This completes the proof. □

<u>Remark 9.2</u>. Babuska and Aziz [1973], Fix [1973], and Kolata

[1978] proved the estimate

$$(9.55) \qquad \lambda_{k_i+j-1,h} - \lambda_{k_i+j-1} \leq C\varepsilon_h^2(\lambda_i), \quad j = 1, \ldots, q_i.$$

where $\varepsilon_h(\lambda_i)$ is defined in (8.21). (9.55) is weaker than (9.8).

For $j = 1, \ldots, q_i - 1$, (9.8) shows higher rates of convergence for

certain problems; see the discussion of multiple eigenvalues in

Subsection 10B. Birkhoff, De Boor, Swartz, and Wendroff [1966]

estimated $\lambda_{k_i+j-1,h} - \lambda_{k_i+j-1}$ in terms of the sum of the squares

of the a-norm distances between S_h and the unit eigenvectors

associated with all the eigenvalues λ_ℓ not exceeding λ_{k_i}.

CHAPTER III. APPLICATIONS

In this chapter we apply the abstract results developed in Chapter II to several representative problems.

Section 10. The Ritz Method for Second Order Problems

A. Vibrations of a Free L-shaped Panel

We consider the problems of the plane strain vibration of an L-shaped panel Ω with free boundary. The specific shape of the panel is shown in Figure 10.1.
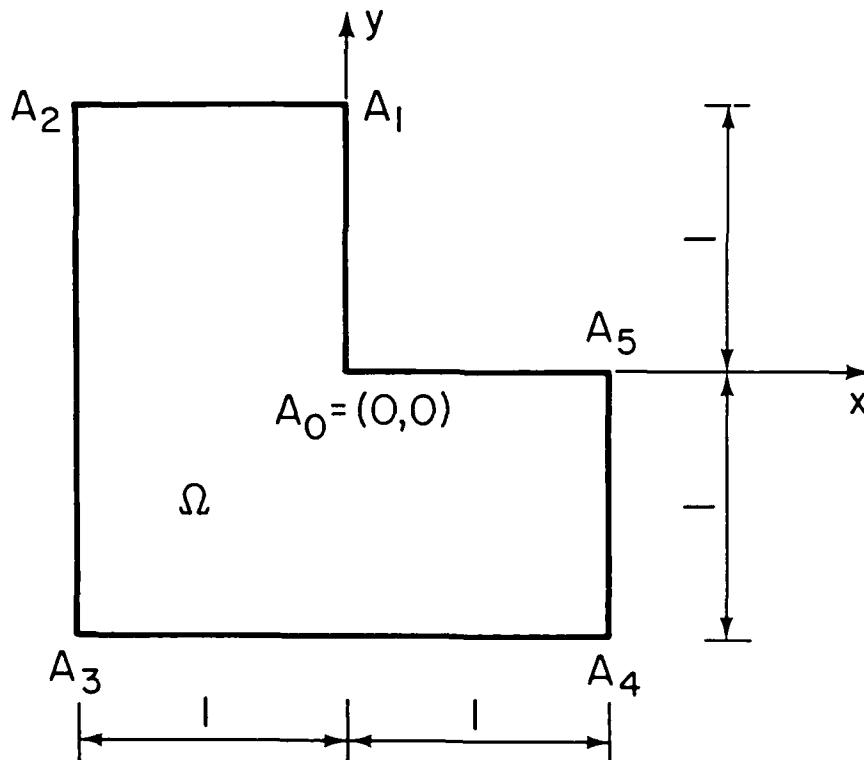

Figure 10.1   The   L-Shaped Panel Ω

The equations governing the vibration of an elastic solid were discussed in Section 1 (see (1.33) - (1.35)). Corresponding to the L-shaped panel we have the eigenvalue problem

111

$$(10.1) \qquad \begin{cases} -(\lambda + \mu)\dfrac{\partial \theta}{\partial x} - \mu \Delta u = \omega \rho u \\[2mm] -(\lambda + \mu)\dfrac{\partial \theta}{\partial y} - \mu \Delta v = \omega \rho v, \quad (x,y) \in \Omega, \end{cases}$$

where $\theta = \dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}$. We obtain (10.1) from (1.35) by assuming that $u(x,y,z)$ and $v(x,y,z)$ are independent of z and that $w(x,y,z) = 0$. The boundary conditions describing the traction free boundary are

$$(10.2) \qquad X_n = Y_n = 0, \quad (x,y) \in \Gamma = \partial \Omega,$$

where

$$(10.3) \qquad \begin{cases} X_n = \lambda \theta n_x + \mu \dfrac{\partial u}{\partial n} + \mu \left(\dfrac{\partial u}{\partial x} n_x + \dfrac{\partial v}{\partial x} n_y\right) \\[3mm] Y_n = \lambda \theta n_y + \mu \dfrac{\partial v}{\partial n} + \mu \left(\dfrac{\partial u}{\partial y} n_x + \dfrac{\partial v}{\partial y} n_y\right) \end{cases}$$

(10.2), with X_n and Y_n given in (10.3) are the Neumann conditions discussed in connection with the elastic solid specialized to the L-shaped panel.

We now consider the specific case in which

$$\nu = \frac{\lambda}{2(\lambda+\mu)} = .3, \quad E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu} = 1$$

(i.e., in which $\lambda = 15/26$ and $\mu = 5/13$). ν is called Poisson's ratio and E is called Young's modulus of elasticity. $G = \mu$ is called the modulus of rigidity. Note that $0 \leq \nu < 1/2$ for any material.

We now discuss the basic steps in the finite element approximation of the eigenvalues and eigenfunctions of the problem (10.1), (10.2), or, more generally, of any eigenvalue problem. These steps are as follows:

1) Derivation of a variational formulation (8.10) for

   (10.1), (10.2) and verification of conditions (8.1) –

112

(8.3), and (8.6) ((8.1),(8.7),(8.32), and (8.33) in the selfadjoint, positive definite case).

2) Discretization of (8.10) and assessment of the accuracy of the approximate eigenvalues and eigenfunctions. The discretization proceeds by the selection of the trial space $S_{1,h}$ and test space $S_{2,h}$, verification of (8.11) - (8.13), consideration of the finite dimensional eigenvalue problem (8.14), and explicit construction of the matrix eigenvalue problem (8.15). The accuracy of the approximation is assessed by means of the application of the results of Section 8.

3) Solution of the matrix eigenvalue problem (8.15).

The accuracy of the approximation method (8.14) depends in a crucial way on the trial and test spaces $S_{1,h}$ and $S_{2,h}$, and their rational selection is strongly influenced by the available information on the eigenfunctions, typically information regarding their regularity. Thus, also of importance is

1') Analysis of the regularity of the eigenfunctions.

Remark 10.1. The approximation methods we will discuss in this section are referred to as Ritz methods, by which we mean that the eigenvalue problems under consideration are selfadjoint and positive definite and that the test and trial space are equal $(S_{1,h} = S_{2,h} = S_h)$; see the discussion in Section 8.

1) Variational Formulation

We begin by casting our problem in the variational form

(10.4)
$$\begin{cases} u \in H \\ a(u,v) = \omega b(w,v), \ \forall \ v \in H, \end{cases}$$

113

where H is an appropriately chosen Hilbert space and a and b
are appropriately selected bilinear forms. This process was ex-
plained in Section 3. We typically proceed as follows. Multiply-
ing the first equation in (10.1) by ϕ, the second by ψ, adding
the resulting equations together, and integrating over Ω, we
obtain

$$\int_{\Omega} \{[-(\lambda + \mu)\frac{\partial\theta}{\partial x} - \mu\Delta u]\phi + [-(\lambda + \mu)\frac{\partial\theta}{\partial y} - \mu\Delta v]\psi\}dxdy$$

(10.5)
$$= \omega\int_{\Omega} \rho(u\phi + v\psi)dxdy.$$

Now, integration by parts shows that

$$\int_{\Omega} \{[-(\lambda + \mu)\frac{\partial\theta}{\partial x} - \mu\Delta u]\phi + [-(\lambda + \mu)\frac{\partial\theta}{\partial y} - \mu\Delta v]\psi\}dxdy$$

$$= \int_{\Omega} (\lambda + \mu)\theta\frac{\partial\phi}{\partial x}dxdy - \int_{\Gamma} (\lambda + \mu)\theta\phi n_x ds$$

$$+ \int_{\Omega} \mu\nabla u\cdot\nabla\phi\, dxdy - \int_{\Gamma} \mu\frac{\partial u}{\partial n}\phi\, ds$$

$$+ \int_{\Omega} (\lambda + \mu)\theta\frac{\partial\psi}{\partial y}dxdy - \int_{\Gamma} (\lambda + \mu)\theta\psi n_y ds$$

(10.6)
$$+ \int_{\Omega} \mu\nabla v\cdot\nabla\psi\, dxdy - \int_{\Gamma} \mu\frac{\partial v}{\partial n}\psi\, ds$$

$$= \int_{\Omega} \{(\lambda + 2\mu)\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\left(\frac{\partial\phi}{\partial x} + \frac{\partial\psi}{\partial y}\right)$$

$$+ \mu[\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\left(\frac{\partial\phi}{\partial y} + \frac{\partial\psi}{\partial x}\right) - 2\frac{\partial u}{\partial x}\frac{\partial\psi}{\partial y} - 2\frac{\partial v}{\partial y}\frac{\partial\phi}{\partial x}]\}dxdy$$

$$- \int_{\Gamma} (X_n\phi + Y_n\psi)ds.$$

Combining (10.5) and (10.6) we see that if $(\omega,(u,v))$ satisfies

114

(10.1) and (10.2), then

$$\int_\Omega \{(\lambda + 2\mu)\left[\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right]\left[\frac{\partial \phi}{\partial x} + \frac{\partial \psi}{\partial y}\right]$$

$$(10.7) \qquad + \mu[\left[\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right]\left[\frac{\partial \phi}{\partial y} + \frac{\partial \psi}{\partial x}\right] - 2\frac{\partial u}{\partial x}\frac{\partial \psi}{\partial y} - 2\frac{\partial v}{\partial y}\frac{\partial \phi}{\partial x}]\}dxdy$$

$$= w\int_\Omega \rho(u\phi + v\psi)dxdy$$

for all smooth (ϕ,ψ), and, conversely, if (10.7) holds for all smooth (ϕ,ψ), then (10.1) and (10.2) hold, provided u and v are smooth $(u,v \in H^2(\Omega))$.

From (10.7) we see how to choose H, a, and b in (10.4). Let

$$(10.8) \qquad \begin{cases} H = H^1(\Omega) \times H^1(\Omega) \\ \|(u,v)\|_H^2 = \|u\|_{1,\Omega}^2 + \|v\|_{1,\Omega}^2 \end{cases}$$

and on H define the bilinear form

$$a(u,v; \phi,\psi) = \int_\Omega \{(\lambda + 2\mu)\left[\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right]\left[\frac{\partial \phi}{\partial x} + \frac{\partial \psi}{\partial y}\right]$$

$$(10.9)$$

$$+ \mu[\left[\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right]\left[\frac{\partial \phi}{\partial y} + \frac{\partial \psi}{\partial x}\right] - 2\frac{\partial u}{\partial x}\frac{\partial \psi}{\partial y} - 2\frac{\partial v}{\partial y}\frac{\partial \phi}{\partial x}]\}dxdy.$$

It is immediate that (8.1) is satisfied and that a is symmetric. Let us remark that $a(u,v;u,v)$ has the physical meaning of the (double) strain energy and that $\sqrt{a(u,v;u,v)}$ is referred to as the energy norm of (u,v). Recall from Section 8 that b is to be defined on a space $W \supset H$. Let

$$(10.10) \qquad \begin{cases} W = L_2(\Omega) \times L_2(\Omega) \\ \|(u,v)\|_W^2 = \|u\|_{0,\Omega}^2 + \|v\|_{0,\Omega}^2 \end{cases}$$

115

and define

$$(10.11) \qquad b(u,v; \phi,\psi) = \int_\Omega \rho(u\phi + v\psi)dxdy.$$

It is immediate that b is symmetric and satisfies (8.7) and (8.33) and that $H \subset W$, compactly. It remains to consider (8.32). Note that since a and b are symmetric, $H^1(\Omega)$ and $L_2(\Omega)$ may be taken to be real.

We begin by expressing $a(u,v; \phi,\psi)$ in terms of the Poisson ratio ν and the modulus of rigidity G:

$$a(u,v; \phi,\psi) = \frac{2G}{1-2\nu} \int_\Omega \{(1-v)\left[\frac{\partial u}{\partial x}\frac{\partial \phi}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial \psi}{\partial y}\right]$$

$$(10.12) \qquad\qquad + \nu\left[\frac{\partial u}{\partial x}\frac{\partial \psi}{\partial y} + \frac{\partial v}{\partial y}\frac{\partial \phi}{\partial x}\right]$$

$$+ \frac{1-2\nu}{2}\left[\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right]\left[\frac{\partial \phi}{\partial y} + \frac{\partial \psi}{\partial x}\right]\}dxdy.$$

From (10.12) we have

$$a(u,v; u,v) = \frac{2G}{1-2\nu} \int_\Omega \{(1-\nu)[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2]$$

$$(10.13) \qquad\qquad + 2\nu\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + \frac{1-2\nu}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}\right)^2\}dxdy$$

$$\geq \frac{2G}{1-2\nu} \int_\Omega \{(1-2\nu)[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2]$$

$$+ \frac{1-2\nu}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2\}dxdy.$$

Recalling that $\nu < 1/2$, we see from (10.13) that

$$a(u,v; u,v) \geq 0, \quad \forall \ u,v$$

(as was to be expected from the physical interpretation), and that $a(u,v; u,v) = 0$ if and only if

116

$$(10.14) \quad \begin{cases} u = \bar{u}_{c_1,c_2,c_3} = c_1 + c_2 y \\ v = \bar{v}_{c_1,c_2,c_3} = c_3 - c_2 x \end{cases}$$

for some c_1, c_2, c_3. These displacements, which are characterized as having no strain energy, are the "rigid body motions," i.e., translations and rotations. Thus (8.32) does not hold with H defined by (10.8), but the above considerations suggest that it might hold if H is replaced by a smaller space that did not include the rigid body motions. In fact, if we define

$$(10.15) \quad \tilde{H} = \{(u,v) \in H : \int_\Omega \rho(u\bar{u}_{c_1,c_2,c_3} + v\bar{v}_{c_1,c_2,c_3})\, dxdy = 0,$$

$$\forall\, c_1,c_2,c_3\},$$

then it can be shown (see Nečas and Hlaváček [1970] and Knops and Payne [1971]) that

$$(10.16) \quad a(u,v;\, u,v) \geq \alpha\|(u,v)\|_H^2 = \alpha(\|u\|_{1,\Omega}^2 + \|v\|_{1,\Omega}^2),$$

$$\forall\, (u,v) \in \tilde{H},$$

where α is a positive constant. This is (8.32).

We thus restrict $a(u,v;\, \phi,\psi)$ to $\tilde{H}$ and $b(u,v;\, \phi,\psi)$ to

$$(10.17) \quad \tilde{W} = \{(u,v) \in W : \int_\Omega \rho(u\bar{u}_{c_1,c_2,c_3} + v\bar{v}_{c_1,c_2,c_3})\, dxdy = 0$$

$$\forall\, c_1,c_2,c_3\}.$$

For the eigenvalue problem (10.1), (10.2) we therefore have the variational formulation

$$(10.18) \quad \begin{cases} 0 \neq (u,v) \in \tilde{H} \\ a(u,v;\, \phi,\psi) = \omega b(u,v;\, \phi,\psi),\ \forall\, (\phi,\psi) \in \tilde{H}. \end{cases}$$

Thus, with a, b, H, $\tilde{H}$, ω, and $\tilde{\omega}$ chosen as in (10.8) - (10.11), (10.15) - (10.17), we see that a and b are symmetric and conditions (8.1), (8.7), (8.32), and (8.33) are satisfied. (10.18) is a selfadjoint, positive definite problem of the type studied at the end of Section 8.

As stated in Section 8, (10.18) has a countable sequence of eigenvalues

$$0 < \omega_1 \le \omega_2 \le \ldots \nearrow +\infty$$

and corresponding eigenfunctions

$$(u_1, v_1), (u_2, v_2), \ldots,$$

which can be chosen to satisfy

$$a(u_i, v_i; u_j, v_j) = \omega_j b(u_i, v_i; u_j, v_j) = \omega_j \delta_{ij}.$$

When implementing our approximation method it is simpler to consider the eigenvalue problem on the space H instead of on $\tilde{H}$, i.e., to consider (10.18) with $\tilde{H}$ replaced by H. Then $\omega_0 = 0$ will be a triple eigenvalue with eigenfunctions (u,v) given by (1,0), (0,1), and (-y,x). These eigenpairs and their approximations are then ignored. If the rigid body motions are included in the space $S_{1,h}$ and $S_{2,h}$, then $\omega_0 = 0$ is also a triple approximate eigenvalue with the rigid body motions again the corresponding approximate eigenfunctions. If this is not the case, then dealing with $\tilde{H}$ and H does not lead to the same approximate eigenvalues and eigenfunctions. It is easy to analyze the case in which the rigid body motions are not in $S_{1,h}$ and $S_{2,h}$, but we will not do so. Alternately, the validity of (10.16) or (8.32) can be ensured by considering

118

$$\tilde{a}(u,v;\, \phi,\psi) = a(u,v;\, \phi,v) + b(u,v;\, \phi,\psi)$$

instead of $a(u,v;\, \phi,\psi)$. Then the triple eigenvalue 1 would be the lowest eigenvalue. Usually the first alternative is used.

1') <u>Regularity of the Eigenfunctions</u>.

We have seen in Section 8 that the accuracy of the approximate eigenvalues and eigenfunctions depends on the degree to which the exact eigenfunctions and adjoint eigenfunctions can be approxmated by elements in the spaces $S_{1,h}$ and $S_{2,h}$, respectively (see (8.23) - (8.26)). In the selfadjoint, positive definite case this reduces to the degree to which the eigenfunctions can be approxi- mated by S_h (see (8.44) - (8.46)). Since the approximability of the eigenfunctions depends on their regularity, it is essential to determine the basic regularity properties of the eigenfunctions.

The eigenfunctions (u_i, v_i) of (10.18) have the following properties.

- u_i and v_i are analytic in $\bar{\Omega} - UA_j$, where A_i are the vertices of Ω. This follows from the general theory of elliptic equations (see Morrey [1966, Section 6.6]).

- The functions u_i and v_i are singular at the vertices of A_j, and the character of the singularity is known. The strength of the singularity at A_j depends on the interior angle at A_j. For the domain we are consider- ing, the strongest singularity is at the vertex A_0. The leading terms of u_i and v_i at a vertex have the form

(10.19)
$$\begin{cases} u^* = C_1 r^\sigma F_1(\theta) \\ v^* = C_2 r^\sigma F_2(\theta), \end{cases}$$

119

where (r,θ) are the polar coordinates with origin at the vertex, σ depends on the interior angle and on λ and μ, and $F_1(\theta)$ and $F_2(\theta)$ are analytic functions of θ. The value of σ is characterized as the root of a nonlinear equation and, in general, can be real or complex. For our example of the L-shaped domain, $\sigma = \sigma_0 = .544481\cdots$ for the vertex A_0. For a more complete discussion of the singularities of solutions of elliptic equations in polygonal domains we refer to Kondratev [1968], Merigot [1974], and Grisvard [1985]. Using their results, any eigenfunction can be written as $(u,v) = (u^1,v^1)+(u^2,v^2)$, where $u^2, v^2 \in H^k(\Omega)$, where k is an integer which is greater than or equal to 3, and (u^1,v^1) is a linear combination of functions of the type on the right side of (10.19) with $\sigma \geq \sigma_0$ and with coordinates centered in the vertices of Ω. Application of the method used in the proof of Theorem 2.1 shows that, for our domain, u^* and v^* and thus u_i and v_i are contained in $H^{\sigma_0+1}(\Omega)$, with $\sigma_0 = .544481\cdots$. This statement of the regularity or smoothness of the eigenfunctions is the strongest that can be made in terms of Sobolev spaces (without weights).

It is also possible to show that $u,v \in \mathcal{B}_\beta^2(\Omega)$, for any $\beta > \sigma_0$ ($\beta = \sigma_0 + \varepsilon$); see Theorem 4.4. Of course, the space $\mathcal{B}_\beta^2(\Omega)$ is much smaller than $H^{\sigma_0+1}(\Omega)$ and hence we can make a more effective choice for S_h if we use $\mathcal{B}_\beta^2(\Omega)$ instead of $H^{\sigma_0+1}(\Omega)$.

2) Discretization of (8.11) and Assessment of the Accuracy of the Approximate Eigenvalues and Eigenfunctions

The discretization of (8.11) is accomplished by selecting the trial and test spaces $S_{1,h}$ and $S_{2,h}$ satisfying (8.11) - (8.13), considering the finite dimensional eigenvalue problem (8.14), and deriving the matrix eigenvalue problem (8.15) from which the approximate eigenvalues are obtained. The selection of $S_{1,h}$ and $S_{2,h}$ is the most important part of this process. It is influenced by three considerations.

a. The spaces $S_{1,h}$ and $S_{2,h}$ have to satisfy (8.11) and (8.12). Note, however, that if the problem under consideration is selfadjoint and positive definite, from (8.32) we see that (8.11) and (8.12) hold for $S_{1,h} = S_{2,h} = S_h$, for any S_h. Our problem (10.18) is self-ajoint and positive definite and we will take $S_{1,h} = S_{2,h} = S_h$.

b. $S_{1,h}$ should accurately approximate the eigenfunctions of (10.18) and $S_{2,h}$ should accurately approximate the adjoint eigenfunctions. Usually we also require that the rigid body motion functions are included in $S_{1,h}$ and $S_{2,h}$. If this is not the case, then we have to assume that the rigid body motion functions are very well approximated. If they are not well approximated, although there will be no change in the asymptotic rate of convergence, the accuracy will deteriorate, especially with long domains (such as long beams), for which the rigid body motions for some parts of the domain could be relatively large.

c. The matrices A and B in (8.15) should be reasonably sparse, since sparcity is strongly related to computational complexity. Sparcity is achieved by choosing finite element spaces for $S_{1,h}$ and $S_{2,h}$. These spaces then have bases consisting of functions with local supports, and, as a consequence, A and B will be sparse. We note that the sparseness of A and B is not required for the validity of the results of Section 8 and, in fact, in certain applications one does use non finite element type trial and test spaces, spaces consisting of global polynomials or trigonometric polynomials, e.g.

We now describe some typical choices for S_h for the L-shaped panel.

The h-Version on a Uniform Mesh

Let Ω be covered by a mesh of uniform squares I_{ij} of size h as shown in Figure 10.2.

122
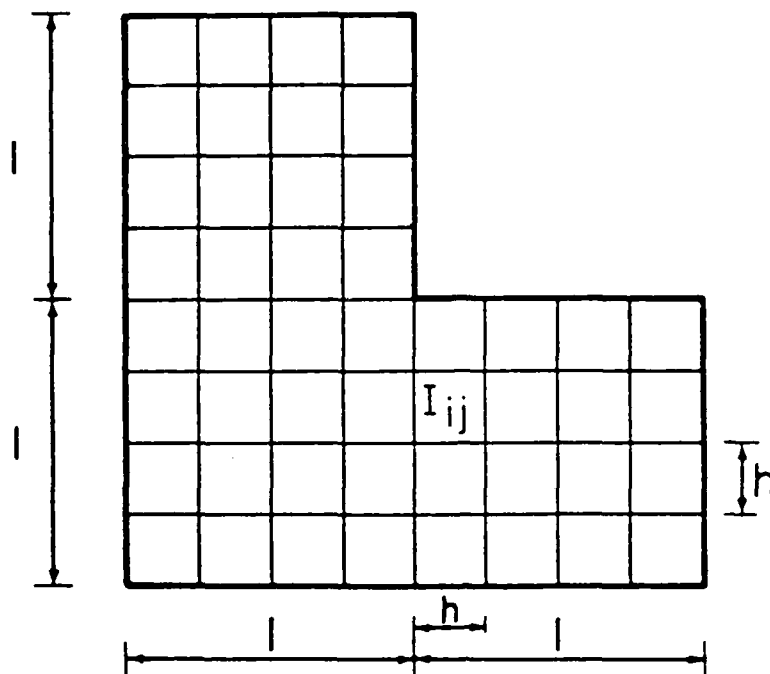
Figure 10.2. A Uniform Mesh on Ω

Then for $p = 1, 2, \ldots$ let

(10.20) $\tilde{S}_h^p = \{u : u \in H^1(\Omega), \; u|_{I_{ij}} = \displaystyle\sum_{(m,n) \in \mathcal{D}(p)} A_{m,n} x^m y^n, \; \forall \; I_{ij}\},$

where

(10.21) $\mathcal{D}(p) = \{(m,n) : 0 \le m, n, \; \text{and} \; m+n \le p \; \text{or} \; (m,n) = (1,p)$

$\text{or} \; (p,1)\}.$

Spaces of this type are said to be composed of elements of type Q'_p (the cases $p = 1, 2, 3$ are discussed in Ciarlet [1978]). Basis functions for these spaces can be constructed in various ways; for example, by means of Lagrange elements (see Ciarlet [1978]) or by use of hierarchial elements.

Regarding the approximation properties of the family $\{\tilde{S}_h^p\}_{0 < h}$,

it can be shown (see Ciarlet [1978]) that if $k \geq 1$ is an integer, then

(10.22) $\quad \inf\limits_{\chi \in \widetilde{S}_h^p} \|u - \chi\|_{1,\Omega} \leq C(p)h^{\mu-1}\|u\|_{k,\Omega}$, for any $u \in H^k(\Omega)$ and any $h > 0$,

where

(10.23) $\qquad\qquad\qquad \mu = \min(k,p+1)$

and $C(p)$ depends on $p, k,$ and Ω, but is independent of u and h. (10.22) is optimal in the sense that h^μ on the right hand side cannot be replaced by a higher power of h when the mesh is uniform. If k is not an integer, then we have

(10.24) $\qquad\qquad \inf\limits_{\chi \in \widetilde{S}_h^p} \|u - \chi\|_{1,\Omega} \leq C(p)h^{\mu-1}\|u\|_{\widetilde{H}^k(\Omega)}$,

with μ given by (10.23). Note that we have not said anything about the dependence of $C(p)$ on p. The proof in Ciarlet [1978] sugggests that $C(p)$ grows with p, and thus could lead to the conclusion that it is improper to use $p > k-1$. However, this conclusion is not justified because, in fact, $C(p) \leq Cp^{-(k-1)}$; see Babuška and Suri [1987b].

We will now derive (10.24) - (10.23) from (10.22) - (10.23) using the method outlined in Section 2 (cf. (2.12) - (2.16) and Theorem 2.1).

Suppose $m < k < m+1$. Since

$$\|u\|_{\widetilde{H}^k(\Omega)} = \sup\limits_{0 < t < \infty} \{t^{-\theta}K(u,t)\},$$

where $\theta = k-m$, we see that

$$K(u,t) \leq t^\theta\|u\|_{\widetilde{H}^k(\Omega)}.$$

124

Let $\varepsilon > 0$. Then for any $t > 0$ there exist $v_t \in H^m(\Omega)$ and $w_t \in H^{m+1}(\Omega)$ such that $u = v_t + w_t$ and

$$\|v_t\|_{m,\Omega} + t\|w_t\|_{m+1,\Omega} \leq K(u,t) + \varepsilon \leq t^\theta \|u\|_{\hat{H}^k(\Omega)} + \varepsilon.$$

Therefore, using (10.22) - (10.23) we can choose $\chi_{1,t}, \chi_{2,t} \in \tilde{S}_h^p$ so that

$$\|v_t - \chi_{1,t}\|_{1,\Omega} \leq C(p)h^{\mu_1 - 1}\|v_t\|_{m,\Omega} \leq C(p)h^{\mu_1 - 1}(t^\theta \|u\|_{\hat{H}^k(\Omega)} + \varepsilon),$$

where $\mu_1 = \min(m,p+1)$, and

$$\|w_t - \chi_{2,t}\|_{1,\Omega} \leq C(p)h^{\mu_2 - 1}\|w_t\|_{m+1,\Omega} \leq C(p)h^{\mu_2 - 1}(t^{\theta-1}\|u\|_{\hat{H}^k(\Omega)} + \varepsilon),$$

where $\mu_2 = \min(m+1,p+1)$. Letting $\chi_t = \chi_{1,t} + \chi_{2,t}$, we thus have

$$(10.25) \quad \|u - \chi_t\|_{1,\Omega} \leq C(p)(h^{\mu_1 - 1}t^\theta + h^{\mu_2 - 1}t^{\theta-1})(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon),$$

$$\text{for any } t > 0.$$

If $p \geq m$, select $t = h$ in (10.25) to obtain

$$\inf_{\chi \in \tilde{S}_h^p} \|u - \chi\|_{1,\Omega} \leq C(p)h^{\theta + \mu_1 - 1}(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon)$$

$$= C(p)h^{k-1}(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon)$$

$$= C(p)h^{\mu-1}(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon),$$

where $\mu = \min(k,p+1)$. If $p < m$, let $t = 1$ in (10.25) to get

$$\inf_{\chi \in \tilde{S}_h^p} \|u - \chi\|_{1,\Omega} \leq C(p)h^{\mu_1 - 1}(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon)$$

$$= C(p)h^p(\|u\|_{\hat{H}^k(\Omega)} + \varepsilon)$$

125

$$= C(p)h^{\mu-1}(\|u\|_{\tilde{H}^k(\Omega)} + \varepsilon),$$

with $\mu = \min(k,p+1)$. Letting $\varepsilon = \|u\|_{\tilde{H}^k(\Omega)}$ in these estimates

yields (10.24) – (10.23).

Now define

(10.26) $$S_{1,h} = S_{2,h} = S_h = \tilde{S}_h^p \times \tilde{S}_h^p.$$

We remark that the rigid body motions belong to S_h (cf. (10.14)).

Since (10.18) is selfajoint and positive definite and satisfies

(8.32), we see that (8.11), with $\beta(h) = \alpha$, and (8.12) hold.

(10.22) and (10.24) show that S_h accurately approximates the exact

eigenfunctions. Thus (10.22) and (10.24), together with a density

argument, show that (8.13) is satisfied. If an appropriate basis

is chosen for S_h, the matrices A and B in (9.15) can be cal-

culated and they will be sparse. Thus the issues raised above in

a., b., and c. have been addressed.

Now consider the problem (8.14) with this choice for $S_{1,h}$

and $S_{2,h}$ and denote its eigenvalues and eigenfunctions by

$$0 < \omega_{1,h} \le \ldots \le \omega_{N,h}$$

and

$$(u_{1,h}, v_{1,h}), \ldots, (u_{N,h}, v_{N,h}),$$

where $N = \dim S_h$. To assess the accuracy of these approximate

eigenpairs, the results of Section 8 will be applied. All of the

hypotheses for these results have now been shown to be satisfied

for our problem and approximation procedure.

<u>Theorem 10.1</u>. Let $S_{1,h}$ and $S_{2,h}$ be selected as in (10.26).

Suppose ω_k is an eigenvalue of (10.18) with multiplicity q,

i.e., suppose $\omega_{k-1} < \omega_k = \omega_{k+1} = \cdots = \omega_{k+q-1} < \omega_{k+q}$. Then

(10.27) $\quad |\omega_{j,h} - \omega_k| \le C(p)h^{1.088962\cdots}$, $\quad j = k,\ldots,k+q-1$.

If $(w_h,z_h) = (u_{j,h},v_{j,h})$, $j = k,\ldots,k+q-1$, then there is a unit eigenfunction $(u,v) = (u_h,v_h)$ of (10.18) such that

(10.28a) $\quad \|u - w_h\|_{1,\Omega} + \|v - z_h\|_{1,\Omega} \le C(p)h^{.544481\cdots}$,

and if (u,v) is a unit eigenfunction of (10.18) corresponding to ω_k, then there is a unit vector $(w_h,z_h) \in$
$\mathrm{sp}\{(u_{k,h},v_{k,h}),\ldots,(u_{k+q-1,h},v_{k+q-1,h})\}$ such that

(10.28b) $\quad \|u - w_h\|_{1,\Omega} + \|v - z_h\|_{1,\Omega} \le C(p)h^{.544481\cdots}$.

If ω_k is simple, the eigenfunction estimates reduce to

(10.28c) $\quad \|u_{k,h} - u_k\|_{1,\Omega} + \|v_{k,h} - v_k\|_{1,\Omega} \le C(p)h^{.544481\cdots}$.

<u>Proof</u>. We saw in Subsection 1') that u_j and v_j are in H^{σ_0+1}, with $\sigma_0 = .544481\cdots$. Thus from (10.22) - (10.24) we have

$$\varepsilon_h = \sup_{j=k,\ldots,k+q-1} \inf_{\chi=(\chi_1,\chi_2)\in S_h} \|(u_j,v_j) - (\chi_1,\chi_2)\|_{H^1(\Omega)\times H^1(\Omega)}$$
$$\le C(p)h^{\sigma_0}.$$

(10.27) and (10.28) follow from this estimate and (8.44) - (8.46).

$\square$

To show the effectiveness of estimates (10.27) - (10.28) we would have to know the exact eigenfunctions and eigenvalues. Because these are not available we consider instead the quantity

(10.29) $\quad Q(p,h) = \inf_{\chi=(\chi_1,\chi_2)\in S_h} a(u^*-\chi_1,v^*-\chi_1,u^*-\chi_2,v^*-\chi_2)$,

where u^* and v^* are given in (10.19). $Q(p,h)$ can be computed numerically. Figure 10.3 shows the graph of

$$\|e\|_{E,R} = \left[Q(p,h) \big/ a(u^*,v^*;u^*,v^*) \right]^{1/2}$$

as a function of h, for various values of p. $\|e\|_{E,R}$ is the relative error in the energy norm measure of the degree to which (u^*,v^*) can be approximated by functions in S_h. The graph, which is plotted in log-log scale, is a straight line and thus

$$\|E\|_{E,R} = Ch^{\alpha},$$

where α is the slope of the line. We see that the slope is very close to the theoretically predicted $\alpha = .544481\cdots$. Increasing p decreases the constant C but does not affect the slope α.

From an analysis of Figure 10.3 we can draw several conclusions:

• To achieve an accuracy of 5% (respectively, 3%) with elements of degree $p = 1$ we would require N to be about 25,000 (respectively, N to be about 170,000) and with elements of degree 2 we would require N to be about 19,000 (respectively, N to be about 124,000). This shows that a uniform or quasiuniform mesh is completely unacceptable for our problem.

• Because the rate of convergence for eigenvalues is twice that for eigenfunctions, we see that the eigenvalues are much cheaper to compute than the eigenfunctions. Roughly speaking, we see that for eigenvalue calculations the required number of unknowns would be about $N = 160$ (respectively, about $N = 400$) for $p = 1$ and about $N = 140$ (respectively, $N = 350$) for $p = 2$.

· While (10.22) qualitatively characterizes the error beha-
vior, it does not give all the desired quantitative information
because C and $\|u\|_{k,\Omega}$ are not known. More precise quantata-
tive information can be gained only by a posteriori analysis. We
will not, however, be able to pursue this direction. For a survey
of results on a posteriori assessment of the quality of finite
element computations, we refer to Noor and Babuska [1987]. A pos-
teriori error analysis is used also in connection with adaptive
approaches, in which the goal is to let the computer construct the
mesh required to achieve the desired accuracy.
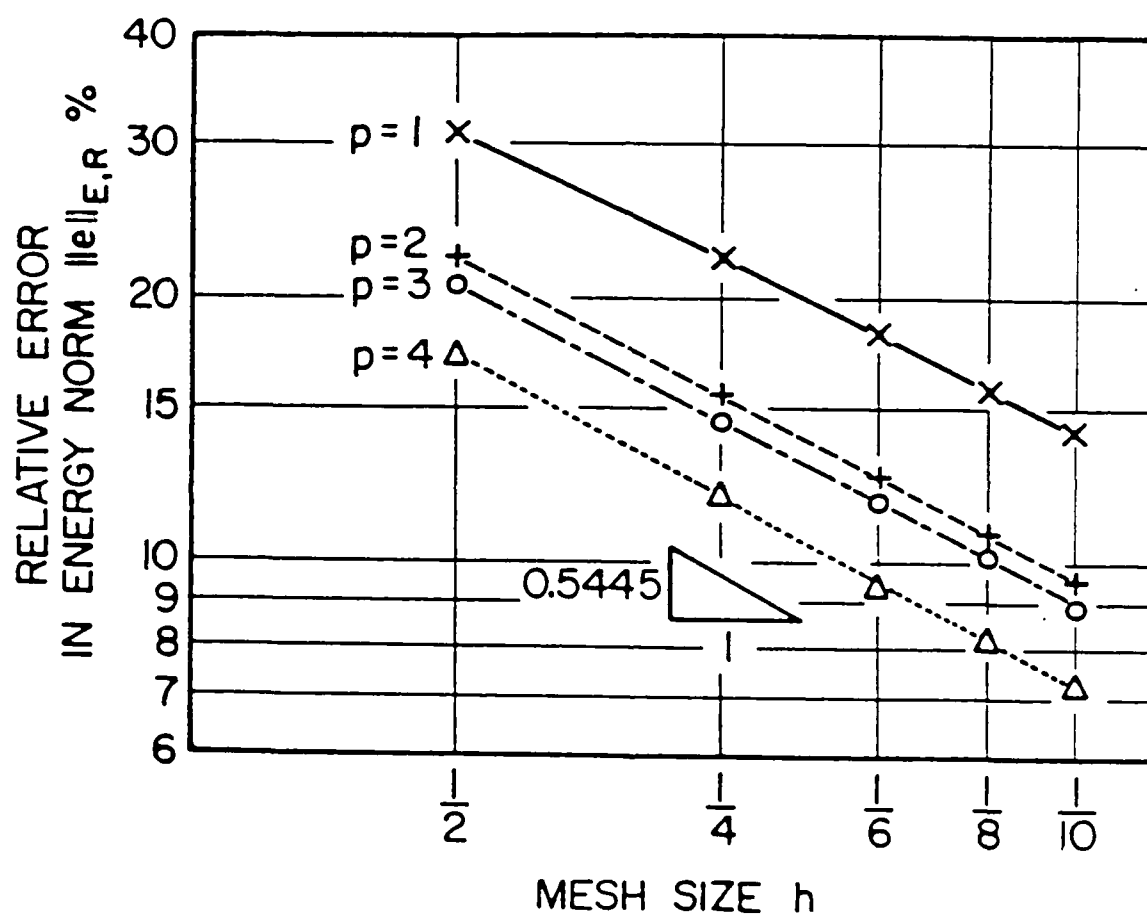
Figure 10.3. The Relative Approximation Error Measured in the Energy Norm. The h-Version.

The p-Version

In the h-version of the finite element method accuracy is achieved by letting $h \searrow 0$, while p is held fixed. In the p-version of the finite element method, one, in contrast, fixes h and lets $p \nearrow \infty$.

Let $\tilde{S}_h^p$ again be defined by (10.20) - (10.21). Regarding the (p-version) approximation properties of the family $(\tilde{S}_h^p)_{p=1,2,\ldots}$, it can be shown that if $u = u_1 + u_2$, where $u_1 \in H^k(\Omega)$, with $k \geq 2$, and $u_2 = Kr^\sigma F(\theta)$, with $\sigma > 0$, then

$$(10.30) \quad \inf_{\chi \in \tilde{S}_h^p} \| u - \chi \|_{1,\Omega} \leq C(h)[Kp^{-2\sigma} + p^{-(k-1)} \| u_1 \|_{k,\Omega}].$$

We remark that in (10.30) it is essential that the origin of Ω lies on an element vertex; for in this case, the estimate for u_2 is of twice the order as would be obtained if we based our estimate on the assumption that $u_2 \in \overset{\approx}{H}{}^{\sigma+1}$ and used the h-version with a uniform mesh. For a proof of (10.30), see Babuška and Suri [1987a].

Define

$$(10.31) \quad S_{1,p} = S_{2,p} = S_p = \tilde{S}_h^p \times \tilde{S}_h^p.$$

Then (8.11), with $\beta(h) = \alpha$, and (8.12) are satisfied. (10.30) shows that S_p accurately approximates the exact eigenfunctions and thus that (8.13) is satisfied. We see that the issues raised in a., b., and c. have been addressed. In connection with c., however, we observe that the matrices A and B are less sparse than with the h-version. Note that the parameter p, which approaches ∞, is here playing the role of the parameter h in

131

Section 8, which approached 0.

Now consider the problem (8.14) with this choice for $S_{1,p}$ and $S_{2,p}$ and denote the eigenvalues and eigenfunctions by

$$0 < \omega_{1,p} \leq \ldots \leq \omega_{N,p}$$

and

$$(u_{1,p}, v_{1,p}), \ldots, (u_{N,p}, v_{N,p}),$$

where $N = \dim S_p$. As with the h-version, the accuracy of the approximate eigenpairs may be assessed with the results of Section 8.

<u>Theorem 10.2.</u> Let $S_{1,p}$ and $S_{2,p}$ be chosen as in (10.31). Suppose ω_k is an eigenvalue of (10.18) with multiplicity q. Then

$$(10.32) \quad |\omega_{j,p} - \omega_k| \leq C(h) p^{-2.177924\cdots}, \quad j = k, \ldots, k+q-1,$$

and

$$(10.33) \quad \|u_{k,p} - u_k\|_{1,\Omega} + \|v_{k,p} - v_k\|_{1,\Omega} \leq C(h) p^{-1.088962\cdots}.$$

Note that we have given the eigenfunction estimate the simplified form it has when ω_k is simple; it would have to be modified if ω_k has multiplicity greater than 1. See the statement of Theorem 10.1.

<u>Proof.</u> Suppose ω_k has multiplicity q and let w be either component of one of the eigenfunctions corresponding to ω_k. We have seen that w can be written in the form $w = w^1 + w^2$, where $w^2 \in H^k(\Omega)$, with $k > 3$, and w^2 is a sum of terms of the type (10.19) with $\sigma > \sigma_0$ and with coordinate centers at the vertices of Ω. Because $\sigma_0 = .544481\cdots$ in (10.19), from (10.30) we have

132

$$\varepsilon_p \leq C(h)p^{-1.088962\cdots}.$$

(10.32) and (10.33) follow from this estimate and (8.44) – (8.46).

□

To illustrate the performance of the p-version we consider, as with the h-version, the relative error in the energy norm measure of the degree to which (u^*, v^*) can be approximated by S_p (cf. (10.29)). Figure 10.4 presents the graph of $\|e\|_{E,R}$ as a function of p, for various values of h. Again the log-log scale is used. We see that the slope is close to the theoretically predicted $-1.088962\cdots$. This is valid only for $p \geq 3$, but recall that all our results are of an asymptotic nature.
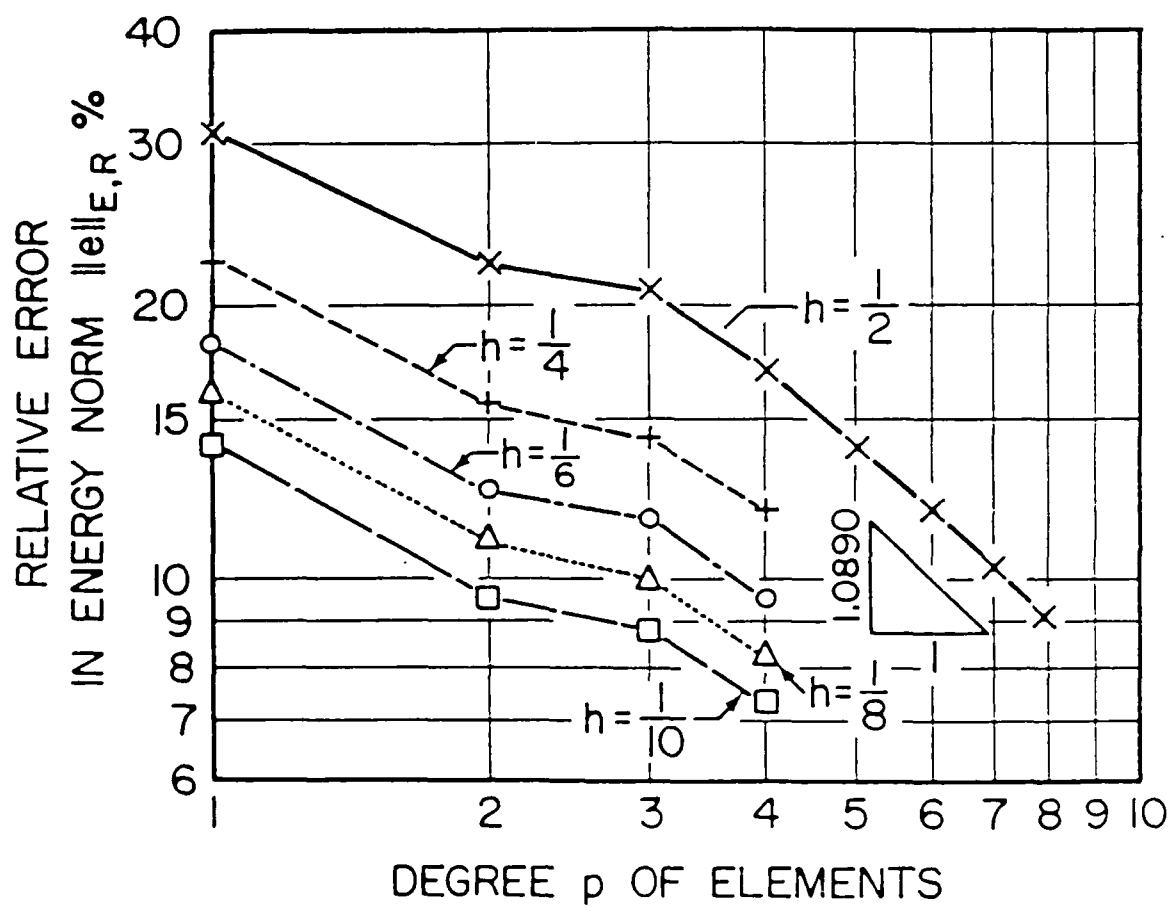
Figure 10.4.   The Relative Approximation Error Measured in the

Energy Norm.   The p-Version.

To assess the relative effectiveness of the h- and p-versions, to understand, in particular, their dependence on the choice of S_h and S_p is not easy. Here we content ourselves with a brief assessment in terms of the number of degrees of freedom: $N = 2 \dim \tilde{S}_h^p = \dim S_h = \dim S_p$. In Figure 10.5, the relative error in the energy norm measure of the accuracy is plotted as a function of N. Since $N \simeq h^{-2}$ and $N \simeq p^2$, the rates of convergence shown in Figure 10.5 are half those shown in Figures 10.3 and 10.4. We see that with respect to degrees of freedom, the p-version with $h = 1/2$ performs better than the h-version with $p = 1,2,3,$ or 4.
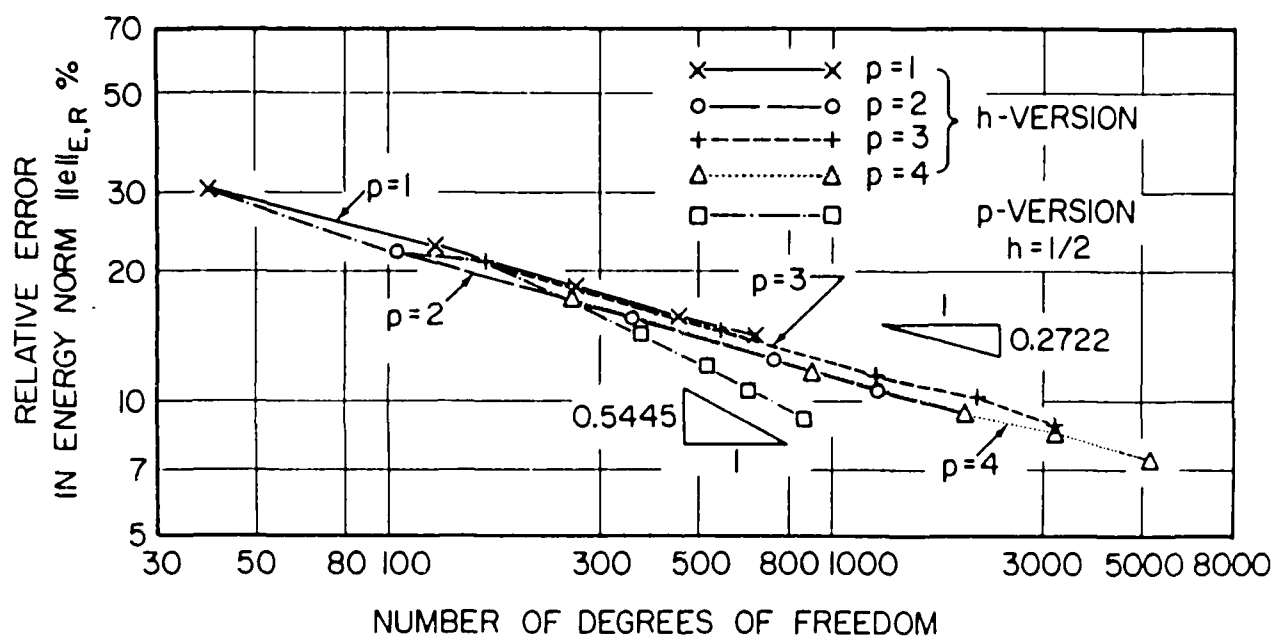
Figure 10.5.   The Relative Error in the Energy Norm in
Dependence on N.

The h-p-<u>Version</u>

In this version of the finite element method accuracy is achieved by simultaneously decreasing the mesh size h and increasing the polynomial degree p. We here distinguish various cases. The major ones are:

i.   Uniform mesh and uniform p distribution, (i.e., the polynomial degree p is the same on each mesh subdomain, i.e., element);

ii.   Refined (non uniform) mesh and uniform p distribution; and

iii. Refined mesh and selective increase of degree p.

We will now elaborate on cases i. and ii.   Case i. obviously combines the h- and p-versions discussed above.   In this case one has

<u>Theorem 10.3</u>. Let $S_{1,(h,p)} = S_{2,(h,p)} = S_{(h,p)} = \tilde{S}_h^p \times \tilde{S}_h^p$ and let $\omega_{k,(h,p)}$ and $(u_{k,(h,p)}, v_{k,(h,p)})$ be the associated eigenvalues and eigenfunctions. Suppose ω_k is an eigenvalue of (10.18) with multiplicity q.   Then

(10.34)
$$|\omega_{j,(h,p)} - \omega_k| \leq C\{\min\left[h^{\sigma_0}, \frac{h^{\min(\sigma_0, p-\sigma_0)}}{p^{2\sigma_0}}\right]\}^2, \quad j = k,\ldots,k+q-1,$$

and

(10.35)
$$\|u_{k,(h,p)} - u_k\|_{1,\Omega} + \|v_{k,(h,p)} - v_k\|_{1,\Omega} \leq C\{\min\left[h^{\sigma_0}, \frac{h^{\min(\sigma_0, p-\sigma_0)}}{p^{2\sigma_0}}\right],$$

where $\sigma_0 = .544481\cdots$ and C is independent of h and p.

<u>Proof</u>. The basic approximation results for this type of approxima-

tion were proved by Babuska and Suri [1987b]. (10.34) and (10.35) follow directly from these results and (8.44) – (8.46). □

In case ii. we will consider only geometric meshes with ratio factor .15; see Figure 10.6. This ratio is close to optimal. The space $S_{(h,p)} = \tilde{S}_h^p \times \tilde{S}_h^p$ is now more complicated. $\tilde{S}_h^p$ is defined by

$$\tilde{S}_h^p = \{u \in H^1(\Omega) : u|_{I_{ij}} \text{ is the image of a polynomial}$$

in a square

$$S = \{(\xi,\eta) : |\xi|, |\eta| \leq 1\}$$

or a triangle

$$T = \{(\xi,\eta) : 0 \leq \eta \leq \xi, \ 0 \leq \xi \leq 1\},$$

for all subdomains I_{ij} in the mesh}.

For a more detailed description of $\tilde{S}_h^p$ see Babuska and Guo [1987b, c] and Szabo [1986]. For a thorough discussion of the h–p–version in the one dimensional setting, we refer to Gui and Babuska [1986].

Figure 10.7 shows the performance of the p–version on meshes with various numbers of layers n, as well as the performance of the p– and the h–versions for uniform meshes. We typically see a reverse S curve for the accuracy of the p–version on a geometric mesh. The first part of the curve is convex and then it is concave, the slope approaching $N^{-\sigma_0}$. The h–p–version appears as the envelope of the p–version on geometric meshes with various numbers of layers. This envelope shows the optimal relation between the number of layers and the polynomial degree. In Babuska and Guo [1987b], it is shown that if $u \in \mathcal{B}_\beta^2(\Omega)$, with $0 < \beta < 1$,

138

then a geometric mesh and a proper selection of the degree p
leads to the exponential rate
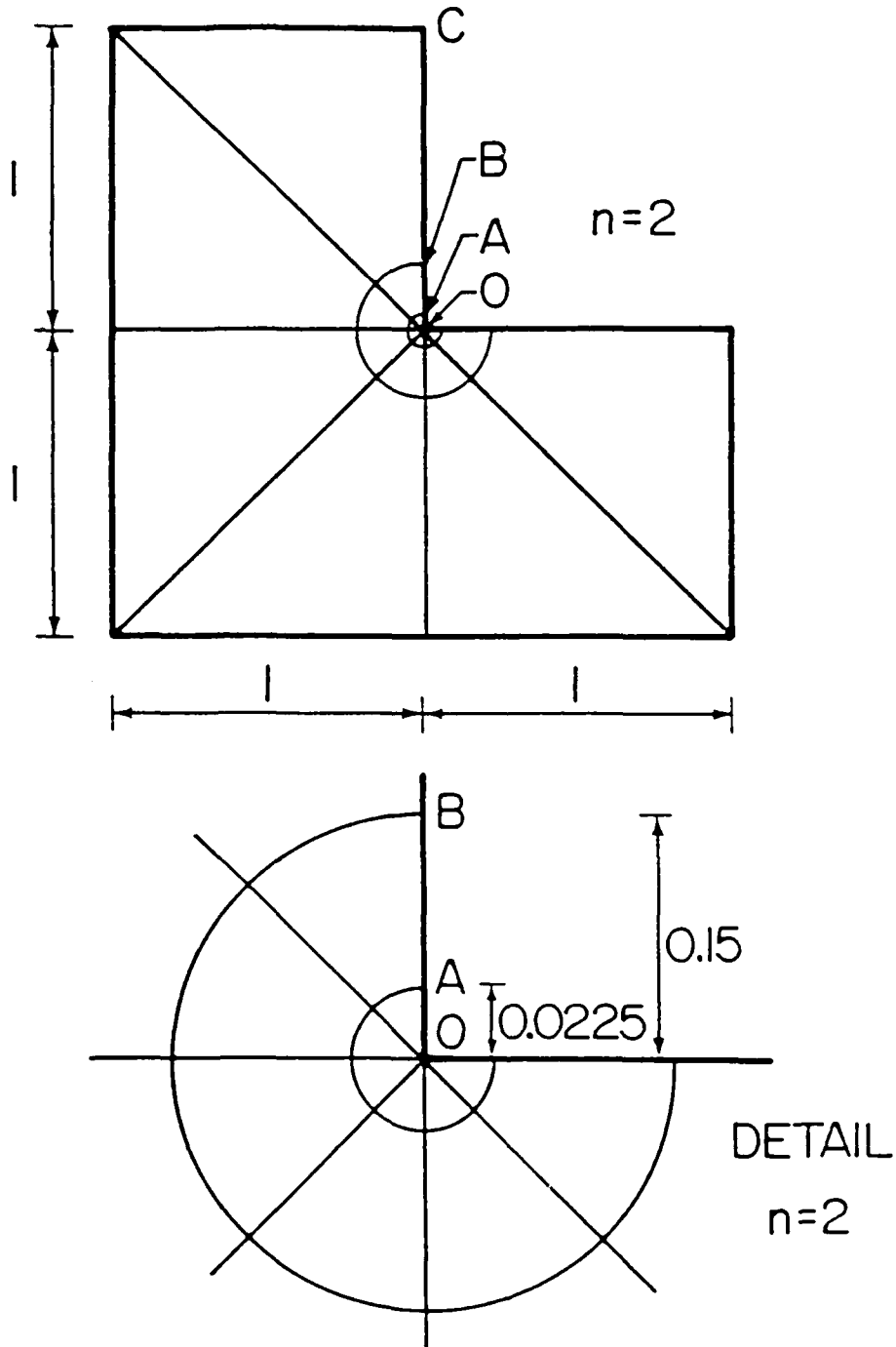
$$\|e\|_{E,R} \le Ce^{-\alpha\sqrt[3]{N}}.$$

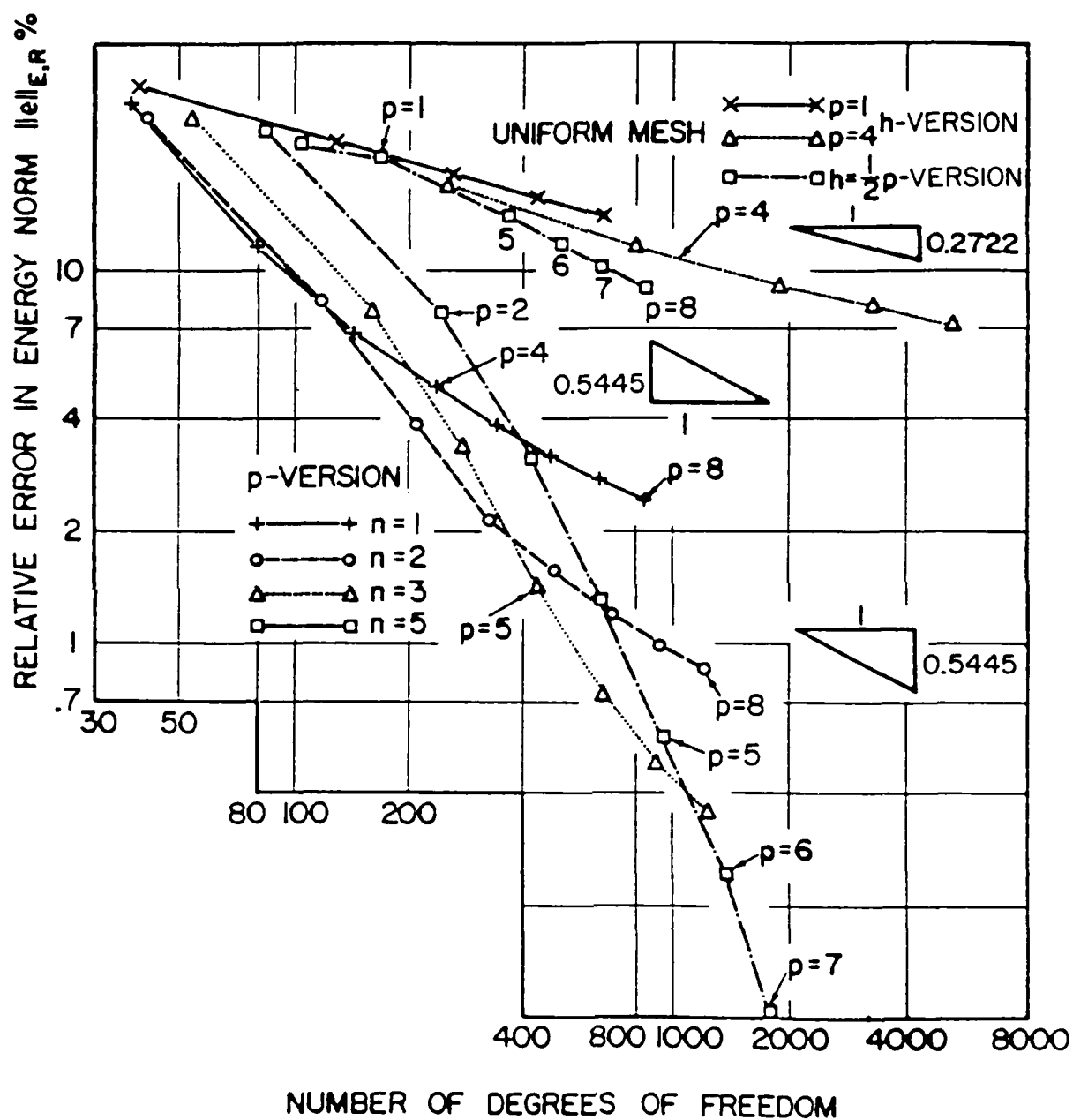Figure 10.6.   The Strongly Refined Mesh with n = 2 Layers.

Figure 10.7. The Relative Error in the Energy Norm in Dependence on N for Various Meshes.

Using this approximation result and the results of Section 8, we obtain

Theorem 10.4. Suppose the components u_i and v_i of the eigenfunctions belong to $\mathcal{B}_\beta^2$ (in our case $\beta = .544481\cdots + \varepsilon$). Then with a proper choice of geometric mesh and the degree p we have

$$(10.36) \qquad |\omega_{k,p} - \omega_k| \leq Ce^{-2\alpha\sqrt[3]{N}}$$

and

$$(10.37) \qquad \|u_{k,p} - u_k\|_{1,\Omega} + \|v_{k,p} - v_k\|_{1,\Omega} \leq Ce^{-\alpha\sqrt[3]{N}},$$

where α depends on the ratio of the mesh, the relation of p and the number of layers, and the domain, but is independent of N.

Proof. (10.36) and (10.37) follow directly from the results of Section 8 and the above estimate for $\|e\|_{E,R}$. $\quad\square$

Figure 10.7 clearly shows the effect of the proper selection of meshes and element degrees on the effectiveness of the finite element method. It also shows that the optimal choice depends on the required accuracy. The design of the mesh and selection of the degree of the elements is a delicate task. Various approaches to deal with this problem are in the research phase. One promising approach is to apply the principles of artificial intelligence (expert systems). For further details we refer to Babuška and Rank [1987]. Figure 10.7 shows only the dependence of the accuracy on the number of degrees of freedom N. It is also essential to judge the complexity of the method with respect to such factors as number of operations, computer architecture, user's interaction, etc. For a detailed study of computer time, accuracy, and perfor-

141

mance for various numbers of degrees of freedom, we refer to Babuska and Scapolla [1987]. We can see directly from Fig 10.7 that the proper mesh design leading to an accuracy of 5% has 2 layers (the ratio of the sizes of elements is of order 50) and p = 3. For an accuracy of 3%, the optimal p is 3 or 4 and the number of layers is 2 or 3 (which leads to size ratios from 50 to 300). The number of degrees of freedom is 200 - 300 (compared with 25,000 - 170,000 for a uniform mesh and p = 1).

3)   <u>Solution</u> <u>of</u> <u>the</u> <u>Matrix</u> <u>Eigenvalue</u> <u>Problem</u>.

We have seen that the approximation procedure developed in Section 8 leads from the eigenvalue problem (8.10) or (10.18) to the generalized matrix eigenvalue problem (8.15), and that the matrices A and B in (8.15) are sparse if the bases for the trial and test spaces are properly chosen. From the error esti-mates in Section 8 we know that the low eigenvalues of (8.15) (approximately 10% of them) give reasonable approximations to the exact eigenvalues. In fact, we have proved convergence for each fixed eigenvalue, but convergence does not occur for a fixed percentage of the available eigenvalues. If A and B have dimension N, then N approximate eigenvalues are available, but $\omega_{[\alpha N],h} \rightarrow \omega_{[\alpha N]}$ as $N \rightarrow \infty$, where $0 < \alpha < 1$ and $[\alpha N]$ is the integral part of αN. The 10% figure we mentioned above is related to engineering accuracy and practice only. The size of the matrix problem will thus be much larger than the number of eigenvalues we are attempting to calculate. The matrix eigenvalue solver, a crucial component of the complete computational proce-dure, should therefore be designed to effectively find the low

eigenvalues of large, sparse, generalized matrix problems. An appropriate version of the Lanczos algorithm is suitable for this class of matrix problems and is often used in practice. We refer to the monographs by Parlett [1980] and Cullum and Willoughby [1985]. Because the extraction of the eigenvalues is very expensive, various "tricks" are used in engineering practice to reduce the sizes of the matrices under consideration. We will not go further in this direction.

Remark 10.2. It should be emphasized that, generally, the goal of the computation is to find, in addition to the eigenpairs, certain functionals of the eigenfunctions (u,v) — e.g., the stress intensity factors, which are combinations of the derivatives of u and v. We will not pursue this direction since it lies beyond the scope of this article. We refer, e.g., to Babuška and Miller [1984] and Szabo and Babuška [1986].

Remark 10.3. The complete computational resolution of an eigenvalue problem is influenced by a wide range of factors. Some of the most impcrtant of these — the smoothness of the eigenfunctions and the approximation properties of the trial and test space, e.g. — have been discussed in detail. Others — the accuracy of the matrix eigenvalue solver and the relation between the accuracy of the matrix solver and the error $\omega_{k,h} - \omega_k$, e.g. — have not been mentioned or have only been mentioned briefly. While these latter factors are important, we will not be able to pursue them. We also note that the important function of a posteriori analysis of computed data has not been discussed. Likewise we have not discussed any adaptive approaches. For some ideas on the assessment

of the quality of the finite element computations we refer to the survey paper of Noor and Babuška [1987].

B. Vibration of a Membrane

We consider here the eigenvalue problem associated with the small, transverse vibration of a membrane stretched over a bounded region Ω in the plane and fixed along its edge $\Gamma = \partial\Omega$, i.e., the eigenvalue problem

$$(10.38) \qquad \begin{cases} -\Delta u = \lambda u, & \text{in } \Omega \\ u = 0 & \text{on } \Gamma \end{cases}$$

(cf. Subsection 1.B., in particular, (1.27)). We turn now to a discussion of the basic steps 1), 1'), 2), and 3) (cf. Subsection A. above) in the finite element approximation of the eigenpairs of (10.38). The discussion can be brief since these steps are similar for the two problems (10.1) - (10.2) and (10.38), in fact for any eigenvalue problem.

1) Variational Formulation

(10.38) is a special case of problem (3.1) and the variational formulation (3.18) of (3.1) was derived in Section 3. Thus we see that the variational formulation of (10.38) is given by

$$(10.39) \qquad \begin{cases} u \in H_0^1(\Omega), \ u \neq 0 \\ \int_\Omega \left[\dfrac{\partial u}{\partial x}\dfrac{\partial v}{\partial x} + \dfrac{\partial u}{\partial y}\dfrac{\partial v}{\partial y} \right] dxdy = \lambda \int_\Omega uv \ dx, \ \forall \ v \in H_0^1(\Omega). \end{cases}$$

Let

$$a(u,v) = \int_\Omega \left[\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right] dxdy = \int_\Omega \nabla u \cdot \nabla v \ dxdy$$

144

be defined for $u, v \in H = H_0^1(\Omega)$, and let

$$b(u,v) = \int_\Omega uv \, dx$$

be defined for $u, v \in W = L_2(\Omega)$. Then (10.39) has the form of (8.10), and a and b are symmetric forms, (8.1), (8.7), (8.32), and (8.33) are satisfied, and $H \subset W$, compactly. All of this can be easily seen for the concrete problem we are considering; it also follows from the more general discussion in Section 3. (10.39) is a selfadjoint, positive definite problem. It has eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow +\infty$$

and corresponding eigenfunctions

$$u_1, u_2, \ldots,$$

which can be chosen to satisfy

$$\int_\Omega \nabla u_i \cdot \nabla v_j \, dxdy = \lambda_j \int_\Omega u_i u_j \, dxdy = \delta_{ij}.$$

1') <u>Regularity of the Eigenfunctions</u>

From Theorems 4.1 - 4.4 we obtain the following regularity results for the eigenfunctions u_i of (10.39) (or (10.38)).

- •) For $k \geq 2$, if $\Gamma = \partial\Omega$ is of class C^k, then $u_i \in H^k(\Omega)$.

- •) If Γ is of class C^∞, then $u_i \in C^\infty(\bar{\Omega})$.

- •) If Γ is analytic, then u_i is analytic in $\bar{\Omega}$.

- •) If Ω is a curved polygon with analytic sides and with vertices $A_0, A_1, \ldots,$ then u_i is analytic in $\bar{\Omega} - UA_j$. u_i is singular at the vertices; the strengths of the

145

singularities depend on the interior vertex angles.
Moreover, $u \in \mathcal{B}^2_\beta(\Omega)$ for properly chosen β.

2) Discretization of (8.10) and Assessment of the Accuracy of the Approximate Eigenpairs

Suppose Ω is a polygon. By a triangulation or mesh on $\bar{\Omega}$ we will mean a finite family $\tau = \{T_i\}_{i=1}^{M(\tau)}$ satisfying

- each T_i is a closed triangle,
- $\bar{\Omega} = \bigcup_{i=1}^{M(\tau)}$,
- for any T_i and $T_j \in \tau$, $T_i \cap T_j = \emptyset$ or a common vertex or a common side.

For $0 < \alpha < \pi$, a triangulation τ is said to be α-regular if the minimal angle of every triangle $T \in \tau$ is greater than or equal to α. For any τ, let

$$h = h(\tau) = \max_{i=1,\ldots,M(\tau)} \text{diam } T_i$$

and

$$\underline{h}(\tau) = \min_{i=1,\ldots,M(\tau)} \text{diam } T_i.$$

τ is said to be q-quasiuniform if

$$\frac{h(\tau)}{\underline{h}(\tau)} \leq q.$$

We will often view triangulations $\tau = \tau_h$ as parameterized by $h = h(\tau)$ and consider families $\gamma = \{\tau\} = \{\tau_h\}$ of triangulations that are α-regular. An example of a $\pi/4$-regular, 1-quasiuniform triangulation of the domain $\Omega = \{(x,y) : -1 < x < 1, -1 < y < 1\}$ is shown in Figure 10.8. It is called a uniform triangulation.
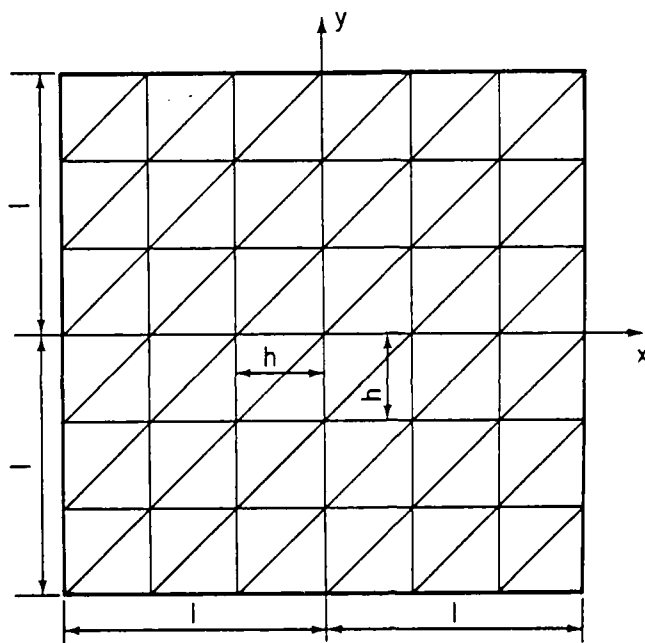
Figure 10.8.   A Uniform Triangulation.

For τ a triangulation of Ω and $p = 1,2,\ldots$ let

$$S^p(\tau) = \{u : u \in H^1(\Omega) : u|_T = \text{a polynomial of degree } p,$$

for each $T \in \tau\}$

and let

$$S_0^p(\tau) = S^p(\tau) \cap H_0^1(\Omega).$$

Regarding the approximation properties of $S^p(\tau)$ and $S_0^p(\tau)$, if $k \geq 1$ (integer or non integer) and $p \geq 1$ and if $\gamma = \{\tau\}$ is a family of α-regular triangulations of Ω, then

(10.40a) $\inf_{\chi \in S^p(\tau)} \|u-\chi\|_{1,\Omega} \leq C \dfrac{h(\tau)^{\mu-1}}{p^{k-1}} \|u\|_{k,\Omega}$,

for any $u \in H^k(\Omega)$ and for any $\tau \in \gamma$,

and

(10.40b) $\inf_{\chi \in S_0^p(\tau)} \|u-\chi\|_{1,\Omega} \leq C \dfrac{h(\tau)^{\mu-1}}{p^{k-1}} \|u\|_{k,\Omega}$,

for any $u \in H^k(\Omega) \cap H_0^1(\Omega)$ and for any $\tau \in \gamma$,

147

where

$$(10.41) \qquad \qquad \mu = \min(k,p+1).$$

The constant C in (10.40a,b) is independent of p, τ, and u, but depends on Ω, k, and α. For a complete proof of these estimates we refer to Babuška and Suri [1987b].

Now define

$$(10.42) \qquad S_{1,(h,p)} = S_{2,(h,p)} = S_{(h,p)} = S_0^p(\tau).$$

Since (10.39) is selfadjoint and positive definite and satisfies (8.32), we see that (8.11) and (8.12) are satisfied. (10.40b) shows that $S_{(h,p)}$ accurately approximates the exact eigenfunctions and thus that (8.13) is satisfied. If a suitable basis is chosen for $S_{(h,p)}$, then the matrices A and B in (8.15) will be sparse. The issues raised in a., b., and c. in Subsection A.2) above have now been addressed for this choice for $S_{(h,p)}$. Note that in using the notation $S_{(h,p)}$ we are identifying $h = h(\tau)$ with τ. An alternate, and more explicit, notation would be $S_{(\tau,p)}$.

Now consider problem (8.14) with $S_{(h,p)}$ defined as in (10.42) and denote the approximate eigenvalues by

$$\lambda_{1,(h,p)} \leq \cdots \leq \lambda_{N,(h,p)}$$

and

$$u_{1,(h,p)} , \cdots , u_{N,(h,p)},$$

where $N = \dim S_{(h,p)}$. To assess the accuracy of these approximate eigenpairs, we apply the results of Section 8, obtaining

Theorem 10.5. Let $S_{(h,p)}$ be selected as in (10.42) and suppose

λ_j is an eigenvalue of (10.39) with multiplicity q. Then

$$(10.43) \qquad |\lambda_{\ell,(h,p)} - \lambda_j| \le C\frac{h^{2\mu-2}}{p^{2k-2}}, \quad \ell = j,\ldots,j+q-1,$$

and

$$(10.44) \qquad \|u_{j,(h,p)} - u_j\|_{1,\Omega} \le C\frac{h^{\mu-1}}{p^{k-1}},$$

where $k \ge 1$ is such that the eigenfunctions corresponding to λ_j are in $H^k(\Omega)$ and $\mu = \min(k,p+1)$. Note that we have given the eigenfunction estimate the simple form it has when λ_j is simple; it would have to be modified in the general case. See the statement of Theorem 10.1.

Proof. Suppose λ_j has multiplicity q. Then (10.39) has eigenfunctions $u_j,\ldots,u_{j+q-1}$, associated with λ_j; by assumption, these eigenfunctions are in $H^k(\Omega)$. Thus, by (10.40) – (10.41), we have

$$(10.45) \qquad \varepsilon_h = \max_{\ell=j,\ldots,j+q-1} \inf_{\chi\in S_{(h,p)}} \|u_\ell - \chi\|_{1,\Omega}$$

$$\le C\frac{h^{\mu-1}}{p^{k-1}} \max_{\ell=j,\ldots,j+q-1} \|u_\ell\|_{k,\Omega} = C\frac{h^{\mu-1}}{p^{k-1}}.$$

(10.43) and (10.44) follow directly from this estimate and (8.44) – (8.46). □

Remark 10.4. If our membrane is free instead of fixed along its edge, then we would have considered the Neumann boundary condition $\frac{\partial u}{\partial n} = 0$. In this situation the eigenvalue problem would have the variational foundation

$$(10.46) \qquad \begin{cases} u \in H, \ u \ne 0 \\ a(u,v) = \lambda b(u,v), \end{cases}$$

where a and b are as above, but

$$H = \{u : u \in H^1(\Omega), \int_\Omega udxdy = 0\},$$

$$\|u\|_H = \|u\|_{1,\Omega}.$$

We would choose

$$S_{(h,p)} = \{u : u \in S^p(\tau), \int_\Omega udxdy = 0\}$$

for the trial and test space. Then all of the hypotheses in Section 8 are satisfied, approximation results similar to (10.40) can be proved, and for the approximate eigenpairs, the error estimates (10.43) and (10.44) follow. We note in particular that the Neumann boundary condition is only implicitly stated in (10.46), i.e., is natural, and thus that the boundary condition need not be imposed on the trial and test functions. This fact makes implementation easier, especially for domains with curved boundaries. See the discussion of natural and essential boundary conditions in Section 3.

3) Solution of the Matrix Eigenvalue Problem

The comments made in Subsection A.3) apply here as well.

Multiple Eigenvalues

The results proved in this Subsection and in Subsection A. cover the case of multiple eigenvalues. Recall that the estimates for $|\lambda_{j,(h,p)} - \lambda_j|$ and $\|u_{j,(h,p)} - u_j\|_{1,\Omega}$ are in terms of

$$\varepsilon_h = \max_{\ell=1,\ldots,q} \inf_{\chi \in S_{1,(h,p)}} \|u_{i_\ell} - \chi\|_{1,\Omega},$$

where q is the multiplicity of λ_j and $u_{i_1},\ldots,u_{i_\ell}$ are the

corresponding eigenfunctions. We now make some comments on multiple eigenvalues and then make an application of the refined error estimates for multiple eigenvalues proved in Section 9.

The eigenvalues and eigenfunctions of the membrane problem on a square, i.e., the problem

(10.47)
$$\begin{cases} -\Delta u = \lambda u & \text{on } \Omega \\ u = 0 & \text{on } \Gamma, \end{cases}$$

where

$$\Omega = \{(x,y) : |x|, |y| < \pi\},$$

are easily seen to be given by

$$\lambda_{k,\ell} = k^2 + \ell^2$$

and

$$u_{k,\ell} = \sin kx \sin \ell y, \quad k, \ell = 1, 2, \ldots .$$

Hence we see that there are multiple eigenvalues. (10.47) is typical of problems with symmetries ((10.47) is symmetric with respect to x and y), and we thus see that multiple eigenvalues are common in applications.

For $i = 1, 2, \ldots,$ let λ_{k_i} be an eigenvalues of (10.47) of multiplicity q_i, i.e., suppose

$$\lambda_{k_i-1} < \lambda_{k_i} = \lambda_{k_i+1} = \cdots = \lambda_{k_i+q_i-1} < \lambda_{k_i+q_i} < \lambda_{k_i+1}.$$

Note that we are here using the notation introduced in Section 10, whereby $k_1 = 1$, k_2 is the lowest index of the second distinct eigenvalue, etc. Suppose now that $q_i > 1$, i.e., that λ_{k_i} is multiple. Let $\{S_h\}$ be any family of finite dimensional subspaces of $H_0^1(\Omega)$ satisfying (9.14). Recall from Section 7 that

151

the q_i approximate eigenvalues

$$\lambda_{k_i,h}, \cdots, \lambda_{k_i+q_i-1,h}$$

converge to λ_{k_i}. While these approximate eigenvalues may be
equal, i.e., we may have one distinct eigenvalue with multiplicity
q_i, consideration of the situation in which we choose S_h to be
$S^1(\tau)$, where τ is the triangulation shown in Figure 10.8., shows
that they may not be equal, since some of the symmetries present
in (10.47) are not present in the discrete problem. Nevertheless,
Theorem 10.5 provides estimates for each of the errors
$|\lambda_{k_i+j-1,h}, \cdots, \lambda_{k_i+j-1}|$, $j = 1,\ldots,q_i$. As we have seen the esti-
mates are

$$|\lambda_{k_i+j-1,h}-\lambda_{k_i+j-1}| \leq C\varepsilon_h^2$$

(10.48)
$$= C[\sup_{u \in M(\lambda_{k_i})} \inf_{\chi \in S^1(\tau)} \|u-\chi\|_{1,\Omega}]^2,$$

$$j = 1,\ldots,q_i,$$

which suggest that the error in λ_{h,k_i+j-1} depends on the degree
to which $S^1(\tau)$ can approximate all of the eigenfunctions corres-
ponding to λ_{k_i}.

   Recall that in Section 9 (Theorem 9.1) we proved refined
estimates, namely,

(10.49) $$|\lambda_{k_i+j-1,h}-\lambda_{k_i+j-1}| \leq C\varepsilon_{i,j}^2(h), \quad j = 1,\ldots,q_i,$$

where

$$\varepsilon_{i,j}(h) = \inf_{u \in M(\lambda_{k_i})} \inf_{\chi \in S^1(\tau)}$$

152

$$a(u,u_{k_i,h}) = \ldots = a(u,u_{k_i+j-2,h}) = 0.$$

Now for the specific problem (10.47), all eigenfunctions have the same smoothness properties and $S^1(\tau)$ with τ given in Figure 10.8, will approximate them all with the same asymptotic accuracy and (10.48) and (10.49) would each lead to the same estimate in terms of h. The multiplicative constants in the estimates could, however, be different. We further note that there are eigenvalue problems for which the different eigenfunctions corresponding to a multiple eigenvalues have strikingly different approximability properties. For such problems (10.49) would provide a striking improvement over (10.48).

As an example of such a problem, consider

$$(10.50) \quad \begin{cases} -\left[\dfrac{1}{\varphi'(x)}\ u'(x)\right]' = \lambda\,\varphi'(x)u, \ x \in I = (-\pi,\pi), \\[2mm] u(-\pi) = u(\pi), \\[2mm] \left[\dfrac{1}{\varphi'}\ u'\right](-\pi) = \left[\dfrac{1}{\varphi'}\ u'\right](\pi), \end{cases}$$

where

$$\varphi(x) = \pi^{-\alpha}|x|^{1+\alpha}\mathrm{sgn}\ x, \ 0 < \alpha < 1.$$

It is easy to check that the eigenvalues and eigenfunctions are as shown in Table 10.1.

Table 10.1

Eigenvalues and Eigenfunctions of the Eigenvalue Problem (10.50)

i	λ_i	u_i
0	0.0	1
1	1.0	$\cos \varphi(x)$
2	1.0	$\sin \varphi(x)$
3	4.0	$\cos 2\varphi(x)$
4	4.0	$\sin 2\varphi(x)$
$\vdots$	$\vdots$	$\vdots$

We see that $\lambda_1 = \lambda_2$, $\lambda_3 = \lambda_4$, etc.

We cast this problem into the variational form (10.1) by choosing

$$H = \{u(x) \ : \ \|u\| = \left[\int_{-\pi}^{\pi} \frac{(u')^2}{\varphi'} \, dx\right]^{1/2} < \infty, \ u(-\pi) = u(\pi), \ \int_{-\pi}^{\pi} \varphi'u\,dx = 0\},$$

$$a(u,v) = \int_{-\pi}^{\pi} u'v' \frac{1}{\varphi'} \, dx,$$

and

$$b(u,v) = \int_{-\pi}^{\pi} uv\varphi'dx.$$

With these choices (8.10) is equivalent to (10.50), with the understanding that the eigenpair $(\lambda_0, u_0) = (0,1)$ of (10.50) is not present in (8.10). Note that $\|u\| = \|u\|_a$. Let $\|u\|_b = (b(u,u))^{1/2}$. The assumptions made in Section 9 are clearly satisfied. Our approximation is defined by (9.4) with

$$S_{1,h} = S_{2,h} = S_h = \{u \in H \ : \ u \text{ linear on}$$
$$(-\pi+jh, \ -\pi+(j+1)h), \ j = 0,1,\ldots,n-1\},$$

where $h = 2\pi/n$ and n is an even integer.

Now this choice for $\{S_h\}$ it is easily seen that

(10.51)
$$\inf_{\chi \in S_h} \| \cos \varphi(x) - \chi \|_a^2 \simeq Ch^2$$

and

(10.52)
$$\inf_{\chi \in S_h} \| \sin \varphi(x) - \chi \|_a^2 \simeq Ch^{1+\alpha}.$$

Hence from Theorem 10.1 we would expect $\lambda_{1,h}$ and $\lambda_{2,h}$, the two approximate eigenvalues that converge to the double eigenvalue $\lambda_1 = \lambda_2$, to have different convergence rates.

From Tables 10.1 and 10.2 we can find the errors in $\lambda_{h,i}$, $i = 1,2,3,4$, for $\alpha = .4$. These errors are plotted in Figure 10.8 in log-log scale. We clearly see the different rates of convergence, specifically seeing the rates h^2 and $h^{1+\alpha} = h^{1.4}$ for the errors in $\lambda_{i,h}$, for $i = 1,3$ and $i = 2,4$, respectively, as suggested by (10.51) and (10.52). It should be noted that the estimates presented in Theorem 10.1 are of an asymptotic nature in that they provide information only for small h (or large n), i.e., for h (or n) in the asymptotic range. From Figure 10.8 we see that for $\alpha = .4$ we are in the asymptotic range quite quickly, say for $n \geq 16$.

Consider $u_{1,h}$ and $u_{2,h}$, the approximate eigenfunctions corresponding to $\lambda_{1,h}$ and $\lambda_{2,h}$, respectively, normalized by $\| \cdot \|_D = 1$. The results of Section 9 suggest that $u_{1,h}$ should be close to $C \cos \varphi(x)$ and $u_{2,h}$ close to $C \sin \varphi(x)$ (cf. (10.51) and (10.52)), where C is such that $C \sin \varphi(x)$ and $C \cos \varphi(x)$ are normalized by $\| \cdot \|_D = 1$, i.e., $C = \pi^{-1/2}$. To illustrate this point we have computed $c_1^{(i)}$ and $c_2^{(i)}$, $i = 1,2,3,4$, so

155

that

$$
K(i) = \begin{cases} \|u_{i,h} - C_1^{(i)} \cos \varphi(x) - C_2^{(i)} \sin \varphi(x)\|_a, & i = 1,2 \\[2em] \|u_{i,h} - C_1^{(i)} \cos 2\varphi(x) - C_2^{(i)} \sin 2\varphi(x)\|_a, & i = 3,4 \end{cases}
$$

is minimal. We would expect that

(10.53) $C_1^{(2)}, \; C_1^{(4)}, \; C_2^{(1)}, \; C_2^{(3)} \simeq 0$

and

(10.54) $C_1^{(1)} = C_2^{(2)} = C_1^{(3)} = C_2^{(4)} \simeq C = .564189583\ldots$.

Table 10.2 shows some of the results for $\alpha = .4$. We see clearly the results predicted in (10.53) and (10.54). Table 10.2 also shows that $K(1) < K(2)$ and $K(3) < K(4)$, as we would expect.

The last columns in Table 10.2 and Figure 10.8 show that the ratios

$$
\frac{\lambda_{i+1,h} - \lambda_{i+1}}{\lambda_{i,h} - \lambda_i} , \quad i = 1,3,
$$

increase as $h \longrightarrow 0$. This shows that in the whole h-range we considered, the approximate eigenvalues converging to a multiple eigenvalue are well separated.

Table 10.2

Numerical Solution of the Eigenvalue Problem (10.50) for $\alpha = .4$

| n	i	$\lambda_{i,h}$	K(i)		$c_1^{(i)}$		$c_2^{(i)}$		$\dfrac{\lambda_{i+1,h} - \lambda_{i+1}}{\lambda_{i,h} - \lambda_i}$
8	1	1.0716754	.2704	0	.5637791	0	-.1124891	-16	1.5562955
	2	1.1115481	.3423	0	-.4151973	-13	.5636998	0	
	3	5.0394692	.1075	+1	.5558919	0	.1317809	-12	1.1943249
	4	5.2414639	.1191	+1	.5022638	-13	.5516234	0	
16	1	1.0175850	.1329	0	.5641633	0	.1596754	-12	2.0041570
	2	1.0352431	.1881	0	-.8916589	-12	.5641519	0	
	3	4.2691915	.5259	0	.5636643	0	.1124328	-13	1.2575063
	4	4.3385100	.5869	0	-.2689727	-12	.5637697	0	
32	1	1.0043740	.6618	-1	.5641879	0	.6411454	-11	2.6003887
	2	1.0113741	.1067	0	.1323421	-10	.5641830	0	
	3	4.0666055	.2589	0	.5641561	0	.1970954	-10	1.4067517
	4	4.0936974	.3067	0	-.7375504	-10	.5641613	0	
64	1	1.0010921	.3305	-1	.5641895	0	.7729760	-9	3.5190001
	2	1.0038431	.6202	-1	.8670648	-9	.5641883	0	
	3	4.0166006	.1289	0	.5641875	0	.3641341	-10	1.6437659
	4	4.0272875	.1653	0	.1415775	-8	.5641858	0	
128	1	1.0002729	.1651	-1	.5641895	0	.4535626	-7	4.9215830
	2	1.0013431	.3665	-1	.3251219	-7	.5641893	0	
	3	4.0041468	.6440	-1	.5641895	0	.4409247	-7	2.0107071
	4	4.0083380	.9135	-1	-.9705611	-8	.5641890	0	
256	1	1.0000682	.8255	-2	.5641896	0	.8070959	-5	7.0542522
	2	1.0004811	.2193	-1	.7269570	-6	.5641895	0	
	3	4.0010365	.3217	-1	.5641896	0	.6435344	-6	2.5706705
	4	4.0026645	.5162	-1	-.2601000	-6	.5641895	0	
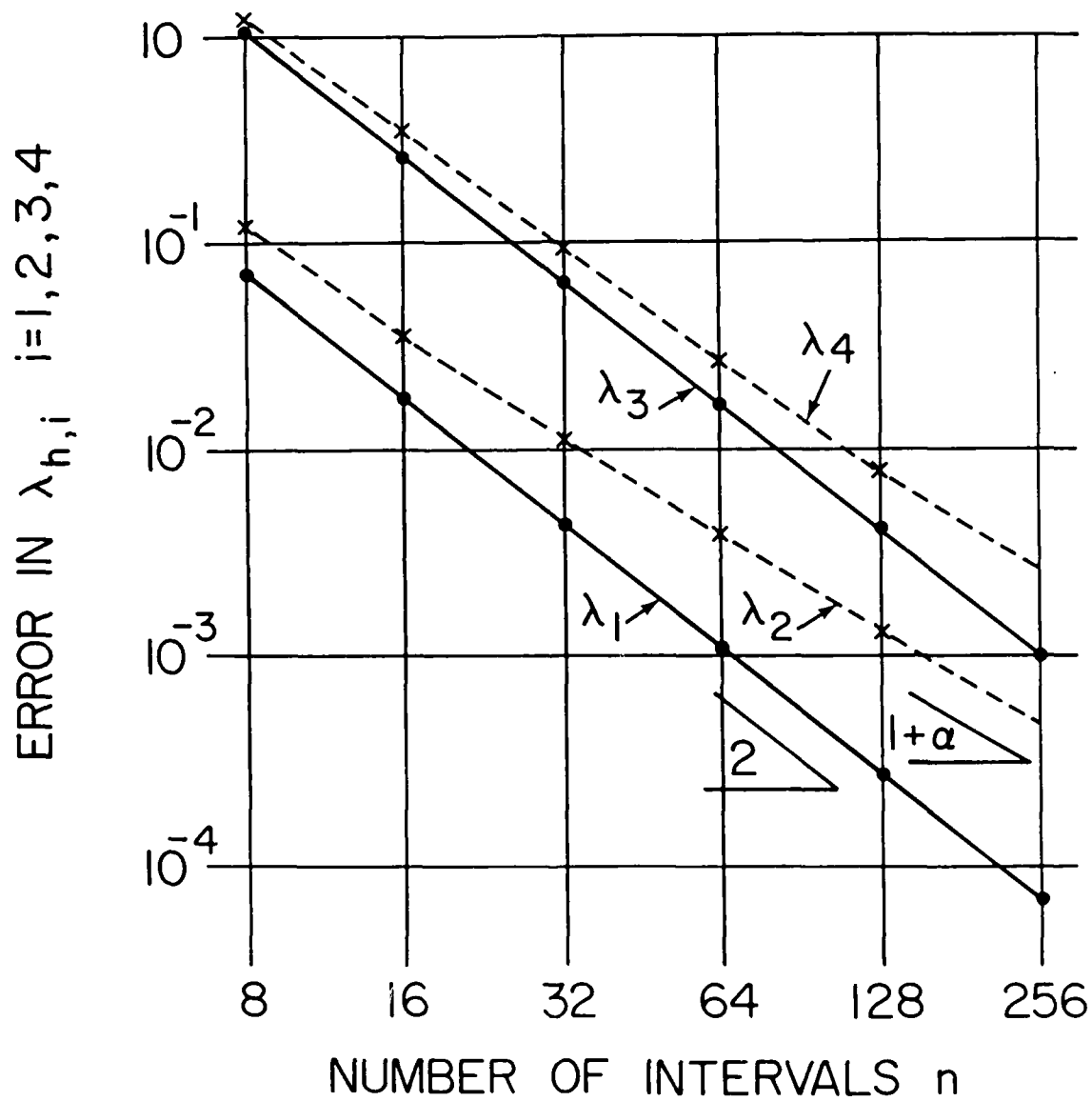
Figure 10.9

The Error in the Eigenvalues $\lambda_{1,h}, \lambda_{2,h}$ and $\lambda_{3,h}, \lambda_{4,h}$ for $\alpha = .4$ in Dependence on the Number of Intervals n

Consider next the case when $\alpha = .01$. Table 10.3 presents
the same results for $\alpha = .01$ as Table 10.2 does for $\alpha = .4$.
Figure 10.9 shows the graph of

$$\log \frac{\lambda_{i+1,h} - \lambda_{i+1}}{\lambda_{i,h} - \lambda_i} , \quad i = 1,3.$$

as a function of the number of intervals n in a semi-logarithmic
scale. The computed values are indicated by 0's and x's. The
graphs are formed by interpolation (solid lines) and extrapolation
(dotted lines). We note three related phenomena that did not occur
with $\alpha = .4$. For small n the approximate eigenfunction associ-
ated with $\lambda_{1,h}$ is $u_{1,h} \simeq \pi^{-1/2} \sin \varphi(x)$, in contrast to $u_{1,h} \simeq$
$\pi^{-1/2} \cos \varphi(x)$ when $\alpha = .4$. We remark that $\pi^{-1/2} \cos \varphi(x)$ is
more easily approximated by S_h than is $\pi^{-1/2} \sin \varphi(x)$ for all
$0 < \alpha < 1$. This anomaly is present for $n \leq 64$ but for $n \geq 128$
we get results which are in agreement with the (asymptotic) results
in Section 9. For $\lambda_{3,h}$ and $\lambda_{4,h}$ we have to take $n \geq 256$ to
get results which agree with the asymptotic theory.

For $\alpha = .01$ we see that $K(2) < K(1)$ for small $n(n \leq 64)$
and $K(2) > K(1)$ for large n and $K(4) < K(3)$ for small
$n(n \leq 128)$ and $K(4) > K(3)$ for large n. Recall that $K(2) >$
$K(1)$ and $K(4) > K(3)$ for all n when $\alpha = .4$.

Finally we note that when $\alpha = .01$ the ratio

$$\frac{\lambda_{i+1,h} - \lambda_{i+1}}{\lambda_{i,h} - \lambda_i}, \quad i = 1,3,$$

first decreases as n increases, then for some n the two eigen-
value errors become equal, and then the ratio increases again.
This is in contrast to the case for $\alpha = .4$, in which the ratio

increased over the whole range of n values. We further note that the value $\bar{n}$ for which the eigenvalue errors are equal — $\bar{n} \simeq 70$ for $i = 1$ and $\bar{n} \simeq 160$ for $i = 2$ (see Figure 10.9) — marks a transition in each of these situations from $u_{1,h} \simeq \pi^{-1/2} \sin \varphi(x)$ to $u_{1,h} \simeq \pi^{-1/2} \cos \varphi(x)$ and $u_{3,h} \simeq \pi^{-1/2} \sin 2\varphi(x)$ to $u_{3,h} \simeq \pi^{-1/2} \cos 2\varphi(x)$, from $K(2) < K(1)$ and $K(4) < K(3)$ to $K(2) > K(1)$ and $K(4) > K(3)$, and from $\dfrac{\lambda_{i+1,h} - \lambda_{i+1}}{\lambda_{i,h} - \lambda_i}$, $i = 1,3$, decreasing to increasing.

We have thus seen that for $\alpha = .4$ the numerical results are in concert with the (asymptotic) results in Section 9 for the whole range of n considered, while for $\alpha = .01$ they are in disagreement for small n, but are in agreement for large n. We now make an observation that further illuminates these two phases of error behavior — the pre-asymptotic and the asymptotic. Toward this end we note that if (λ_1, u_1), with $\|u_1\|_b = 1$, and $(\lambda_{1,h}, u_{1,h})$, with $\|u_{1,h}\|_b = 1$, are first eigenpairs of (9.1) and (9.4), respectively, then

$$0 \le \lambda_{1,h} - \lambda = \|u_{1,h} - u\|_a^2 - \lambda_1 \|u_{1,h} - u_1\|_b^2$$

(10.55)
$$= \inf_{\substack{\chi \in S_h \\ \|\chi\|_b = 1}} [\|\chi - u_1\|_a^2 - \lambda_1 \|\chi - u_1\|_b^2].$$

If λ_1 is a multiple eigenvalue, then the u_1 in (10.56) can be any corresponding eigenvector with $\|u_1\|_b = 1$. (Note that we are here assuming u_1 and $u_{1,h}$ have $\|\cdot\|_b$-length equal 1, whereas in (10.2) and (10.5) they are assumed to have $\|\cdot\|_a$-length equal to 1.) The first inequality in (10.55) follows from the minimum

160

principle (8.35) and has already been stated in (8.42). The first

equality in (10.56) follows immediately from Lemma 9.1 with (λ, u)

$= (\lambda_1, u_1)$, $w = u_{1,h}$, and $\tilde{\lambda} = a(u_{1,h}, u_{1,h}) = \lambda_{1,h}$. If $\chi \in S_h$

with $\|\chi\|_b = 1$, then from the minimum principle (8.35),

$$(10.56) \qquad \lambda_{1,h} - \lambda_1 \leq a(\chi, \chi) - \lambda_1.$$

Again from Lemma 9.1, this time with with $(\lambda, u) = (\lambda_1, u_1)$, $w = \chi$,

and $\tilde{\lambda} = a(\chi, \chi)$, we have

$$(10.57) \qquad a_0(\chi, \chi) - \lambda_1 = \|\chi - u_1\|_a^2 - \lambda_1 \|\chi - u_1\|_b^2.$$

The second equality in (10.55) follows from (10.56) and (10.57).

It is clear from the above discussion that u_1 can be any eigen-

vector corresponding to λ_1.

From (10.55) we have

$$(10.58) \quad \lambda_{1,h} - \lambda_1 \leq \|\chi - u_1\|_a^2 - \lambda_1 \|\chi - u_1\|_b^2, \ \forall \, \chi \in S_h \ \text{ with } \ \|\chi\|_b = 1.$$

If χ is $\|\cdot\|_a$-close to u_1, to be more precise, if χ is taken

to be the a-projection of u_1 onto S_h (cf. (8.17)), then the

second term as the right side of (10.58) is negligible with respect

to the first term. This follows from the compactness assumption

made in Section 9. On the other hand, if $\|u_1 - \chi\|_a$ is not small,

$\lambda_{1,h} - \lambda_1$ may still be small because of cancellation between the

two terms on the right side of (10.58). Regarding the case $\alpha =$

.01, this explains why for h large (the pre-asymptotic phase), we

can have $u_{1,h} \simeq \pi^{-1/2} \sin \varphi \, (x)$ and $K(1) > K(2)$, and yet have

$\lambda_{h,1}$, the approximate eigenvalue associated with $u_{1,h}$, closer to

λ_1 than is $\lambda_{2,h}$, the approximate eigenvalue associated with

$u_{2,h} \simeq \pi^{-1/2} \cos \varphi(x)$, while for h small (the asymptotic phase),

161

we have $u_{1,h} \simeq \pi^{-1/2} \cos \varphi (x)$, $k(1) < k(2)$, and $\lambda_{1,h}$ closer to λ_1 than is $\lambda_{2,h}$, showing that the eigenvalue error, $\lambda_{i,h} - \lambda_i$, is governed by $\inf_{\chi \in S_h} \|\chi - u_1\|_a^2$.

The analysis of example (10.50) we have presented is taken from Babuska and Osborn [1987].

Table 10.3

Numerical Solution of the Eigenvalue Problem (4.1) for $\alpha = .01$

n	i	$\lambda_{h,i}$	K(i)		$c_1^{(i)}$		$c_2^{(i)}$		$\dfrac{\lambda_{h,i+1}-\lambda_{i+1}}{\lambda_{h,i}-\lambda_i}$
8	1	1.0520268	.2338	0	.8181940	-11	.5634386	0	1.0171143
	2	1.0529172	.2268	0	.5645965	0	-.2916448	-11	
	3	4.8576239	.9593	0	-.9346720	-13	.5597529	0	1.0164293
	4	4.8717141	.9615	0	.5604533	0	.1167277	-11	
16	1	1.0128661	.1223	0	.8717399	-10	.5635957	0	1.0111689
	2	1.0130098	.1052	0	.5647369	0	-.8480131	-9	
	3	4.2088367	.4650	0	.2507177	-10	.5636658	0	1.0087030
	4	4.2106542	.4577	0	.5642694	0	-.3101833	-10	
32	1	1.0032139	.7274	-1	-.9345818	-9	.5636031	0	1.0068764
	2	1.0032360	.3568	-1	.5647430	0	.1273043	-7	
	3	4.0515675	.2384	0	.3745461	-9	.5638178	0	1.0057284
	4	4.0518629	.2205	0	.5644172	0	-.4115544	-9	
64	1	1.0008063	.5369	-1	-.1311961	-5	.5636032	0	1.0017363
	2	1.0008077	.3398	-1	.5647430	0	.2462939	-7	
	3	4.0128623	.1343	0	.2743681	-7	.5638240	0	1.0035997
	4	4.0129086	.9792	-1	.5644235	0	.3196172	-8	
128	1	1.0002018	.4196	-1	.5647430	0	.3356056	-5	1.0064420
	2	1.0002031	.4775	-1	.7414162	-6	.5636032	0	
	3	4.0032196	.9166	-1	.2379072	-6	.5638239	0	1.0010560
	4	4.0032230	.9745	-2	.5644235	0	.1197135	-5	
256	1	1.0000504	.4372	-1	.5647429	0	.1061527	-4	1.0218254
	2	1.0000515	.4614	-1	-.1553659	-4	.5636031	0	
	3	4.0008054	.5011	-1	.5644234	0	-.2123278	-4	1.0031040
	4	4.0008079	.7741	-1	.1165012	-5	.5638238	0	
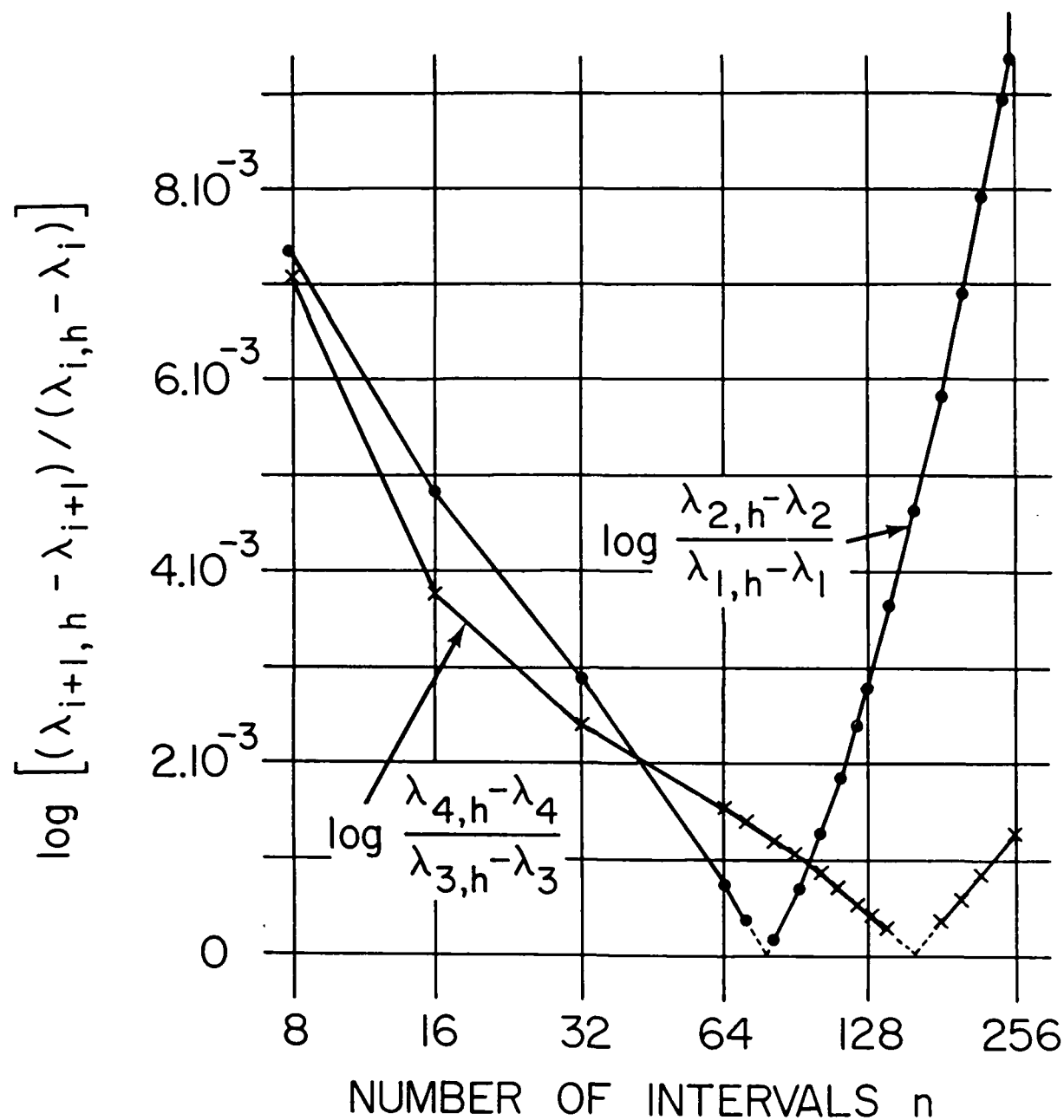
Figure 10.10

The Graphs of $\log \frac{\lambda_{2,h} - \lambda_2}{\lambda_{2,h} - \lambda_2}$ and $\log \frac{\lambda_{4,h} - \lambda_4}{\lambda_{3,h} - \lambda_3}$ for $\sigma = .01$ in

Dependence on the Number of Intervals n.

C. Eigenvalue Problems for General Second Order Elliptic Operators.

We consider here the approximation of the eigenpairs of general second order elliptic operators. This problem is, in large part, similar to those discussed in Subsections A. and B. above; we will thus be brief, discussing in detail only those issues that have a treatment in this case that differs from that for the L-shaped panel or the membrane, or those issues that did not arise with those problems.

Consider the eigenvalue problem

$$(10.59) \qquad \begin{cases} (Lu)(x) = \lambda (Mu)(x), \ x \in \Omega \\ (Bu)(x) = 0, \ x \in \Gamma = \partial\Omega, \end{cases}$$

where Ω is a polygonal domain in R^2, L is given in (3.2), M in (3.3), and B in (3.4), L is assumed to be uniformly strongly elliptic (cf. (3.5)), $a_{ij}, b_i, c,$ and d to be bounded and measurable, and d to be bounded below by a positive constant (cf. Section 2).

In Section 2 we saw that (10.59) has the variational form (8.10), (cf.(3.18) and (3.20)), with $H_1 = H_2 = H_0^1(\Omega)$ in the case of Dirichlet boundary conditions and $H_1 = H_2 = H_0^1(\Omega)$ in the case of Neumann conditions, and a and b given in (3.14). (8.1) - (8.3) hold, (8.2) and (8.3) being a consequence of (3.5), provided

$$(10.60) \qquad \qquad Re \ c(x) \geq \frac{a_0}{2} + \frac{b^2}{2a_0},$$

where $b = \max\limits_{\substack{x \in \Omega \\ i=1,2}} |b_i(x)|$ (cf. 3.17). We remark that (10.60) can be easily achieved. It does not hold for the given operator L, L

165

can be modified, by adding an appropriate multiple of $d(x)$ to $c(x)$, so that it does hold. This change shifts the eigenvalues and leaves the eigenfunctions unchanged. We also see that (8.7) is satisfies with $W_1 = W_2 = L^2(\Omega)$. Thus (10.59) has the form of the problem analyzed in Section 8.

We remark that in this subsection, since we are not imposing any selfadjointness assumptions, the spaces H and $S_{1,(h,p)} = S_{2,(h,p)}$ must be taken to be complex and the eigenvalue parameter λ must be considered complex.

As we have seen, the selection of the trial and test spaces $S_{1,(h,p)}$ and $S_{2,(h,p)}$ is guided by the regularity properties of the exact eigenfunctions and adjoint eigenfunctions. In general, determining this regularity and then using it to choose effective trial and test spaces is a delicate task. The regularity can depend on the coefficients in the differential equation, e.g., on where they have discontinuities and where they are smooth, and on the domain, as we have seen with the L-shaped panel. We will not go further in this direction, but will instead assume the eigen-functions belong to $H^{k_1}(\Omega)$ and the adjoint eigenfunctions to $H^{k_2}(\Omega)$, and select trial and test spaces so as to reflect this assumption.

Remark 10.5. For eigenvalue problems with rough coefficients, which arise in the analysis of vibrations in structures with rapid-ly changing material properties (such as composite materials) it is known that the eigenfunctions do not lie in any high order Sobolev space. Nevertheless, for one dimensional problems, their regular-ity can be understood and, based on this understanding, one can

select trial and test spaces that lead to very accurate and robust
approximations. These trial and test spaces are not of the usual
polynomial type, but instead closely reflect the coefficients.
For details see Babuška and Osborn [1983, 1985]. Cf. also Subsec-
tion 11.C.

Remark 10.6. The mathematical study of the use of regularity
information for the optimal selection of trial and test functions
belongs to the area of complexity and information based approaches.
See, e.g., Wozniakowski [1985].

Based on the information that the eigenfunctions lie in
$H^{k_1}(\Omega)$ and the adjoint eigenfunctions in $H^{k_2}(\Omega)$, with $k_1, k_2 \geq$
1, it is appropriate to discretize (8.10) by choosing

$$
S_{(h,p)} = S_{1,(h,p)} = S_{2,(h,p)} = \begin{cases} S_0^p(\tau), & \text{for Dirichlet conditions} \\ S^p(\tau), & \text{for Neumann conditions,} \end{cases}
$$

as in Subsection B., where $\tau \in \gamma$ and $\gamma = \{\tau\} = \{\tau_h\}$ is a fami-
ly of α-regular triangulations of Ω. (8.11), with $\beta(h) = a_0/2$,
and (8.12) follow from (3.5). (8.13) follows from (10.40).

(8.14) (or (8.15)) can now be considered and from it we get
eigenpairs $(\lambda_{(h,p)}, u_{(h,p)})$ which serve as approximations to the
eigenpairs (λ, u) of (10.59) (or (8.10)). The errors in the
approximate eigenpairs can be estimated with the results of Sec-
tion 8.

Let λ be an eigenvalue of (10.59) (or (8.10)) with alge-
braic multiplicity m (by which we mean that λ^{-1} is an eigen-
value with algebraic multiplicity m of the compact operator T
introduced in (8.8)). Recall that

167

$M(\lambda)$ = the unit ball (with respect to $H^1(\Omega)$) in the space

of generalized eigenfunctions associated with λ,

and

$M^*(\lambda)$ = the unit ball in the space of generalized adjoint

eigenvectors associated with λ.

From (10.40b) in the case of the Dirichlet problem and (10.40a) in the case of the Neumann problem we have

$$\varepsilon_{(h,p)}(\lambda) = \sup_{u \in M(\lambda)} \inf_{\chi \in S_{h,p}} \| u - \chi \|_{1,\Omega}$$

$$\leq C \frac{h^{\mu_1 - 1}}{p^{k_1 - 1}} \sup_{u \in M(\lambda)} \| u \|_{k_1, \Omega},$$

where $\mu_1 = \min(p+1, k_1)$, and

$$\varepsilon^*_{(h,p)}(\lambda) = \sup_{v \in M^*(\lambda)} \inf_{\eta \in S_{h,p}} \| v - \eta \|_{1,\Omega}$$

$$\leq C \frac{h^{\mu_2 - 1}}{p^{k_2 - 1}} \sup_{v \in M^*(\lambda)} \| u \|_{k_2, \Omega},$$

where $\mu_2 = \min(p+1, k_2)$.

Let $\lambda_1(h,p), \ldots, \lambda_m(h,p)$ be the eigenvalues of (8.14) that converges to λ, let

$M_{h,p}(\lambda)$ = {u : u in the direct sum of the generalized eigen-

spaces of (8.14) corresponding to the eigen-

values $\lambda_1(h,p), \ldots, \lambda_m(h,p)$, $\| u \|_{1,\Omega} = 1$},

and let α = ascent of $(\lambda^{-1} - T)$.

Applying Theorem 8.2 we have

$$(10.61) \qquad \left| \lambda - \left(\frac{1}{m} \sum_{j=1}^{m} \lambda_j(h,p)^{-1} \right)^{-1} \right| \leq C \varepsilon_{(h,p)}(\lambda) \varepsilon^*_{(h,p)}(\lambda)$$

$$\le C \frac{h^{\mu_1 + \mu_2 - 2}}{p^{k_1 + k_2 - 2}} \sup_{u \in M(\lambda)} \|u\|_{k_1, M} \sup_{v \in M^*(\lambda)} \|v\|_{k_2, \Omega}.$$

In light of Remark 8.1 we also have

$$|\lambda - \frac{1}{m} \sum_{j=1}^{m} \lambda_j(h, p))|$$

(10.62)

$$\le C \frac{h^{\mu_1 + \mu_2 - 2}}{p^{k_1 + k_2 - 2}} \sup_{u \in M(\lambda)} \|u\|_{k_1, M} \sup_{v \in M^*(\lambda)} \|v\|_{k_2, \Omega}.$$

From Theorem 8.3 we obtain

$$(10.63) \quad |\lambda - \lambda_j(h, p)| \le C \left[\frac{h^{\mu_1 + \mu_2 - 2}}{p^{k_1 + k_2 - 2}} \sup_{u \in M(\lambda)} \|u\|_{k_1, M} \sup_{v \in M^*(\lambda)} \|v\|_{k_2, \Omega} \right]^{1/\alpha}.$$

Regarding eigenfunction estimates, we apply Theorem 8.1 and 8.2. From Theorem 8.1 we have

$$(10.64) \quad \hat{\delta}(M(\lambda), M_{(h,p)}(\lambda)) \le C \frac{h^{\mu_1 - 1}}{p^{k_1 - 1}} \sup_{u \in M(\lambda)} \|u\|_{k_1, \Omega}.$$

Let $\lambda(h, p)$ be an eigenvalue of (10.1) (or (8.10)) such that $\lim_{\substack{h \to 0 \\ p \to \infty}} \lambda(h, p) = \lambda$ and let $w_{(h,p)}$ is a unit vector satisfying $(\lambda(h,p)^{-1} - T)^{\ell_1} w_{(h,p)} = 0$ for some positive integer $\ell_1 \le \alpha$. Then, from Theorem 8.4, for any integer ℓ_2 with $\ell_1 \le \ell_2 \le \alpha$, there is a vector $u_{(h,p)}$ such that $(\lambda^{-1} - T)^{\ell_2} u_{(h,p)} = 0$ and

$$(10.65) \quad \|u_{(h,p)} - w_{(h,p)}\|_{1, \Omega} \le C \left[\frac{h^{\mu_1 - 1}}{p^{k_1 - 1}} \sup_{u \in M(\lambda)} \|u\|_{k_1, \Omega} \right]^{(\ell_2 - \ell_1 + 1)/\alpha}.$$

Remark 10.7. In this section we have considered triangular meshes.

169

One could also consider quadrilateral meshes, which are a general-ization of the type of mesh employed in Subsection A., or curvili-near meshes. Since these generalizations properly belong to approximation theory we will not pursue them. We refer the reader to Ciarlet [1978], Babuśka and Guo [1978b], and Szabo [1986].

Remark 10.8. We have mentioned here only estimates based on the information that $u \in H^k(\Omega)$. If we know, e.g., that $u \in \mathcal{B}_\beta^2(\Omega)$, then we can say more, provided a proper mesh is selected.

Remark 10.9. The approximate eigenvalues $\lambda_j(h,p)$ here, as in any finite element method, are defined by the eigenvalue problem (8.14), which involves integrals over the domain Ω. In practice these integrals often must be evaluated (approximated) by quadra-ture formulas. For estimates of the eigenvalue error due to this quadrature error we refer to Fix [1972]. We note that the use of a finite element method in conjunction with a quadrature method often leads to a finite different method for eigenvalue approxima-tion. For example, if we approximate the eigenvalues of

$$\begin{cases} -\Delta u = \lambda u, & \text{on } \Omega \\ u = 0, & \text{on } \Gamma \end{cases}$$

with the finite element method corresponding to $p = 1$ and a uni-form triangulation (cf. Figure 10.8) and evaluate the resulting integrals with an appropriate quadrature formula, we obtain the standard 5-point difference eigenvalue approximation for the Laplacian (cf. Section 5). This observation is due to Courant [1927, 1943]. For further results on finite difference methods we refer to Polya [1952], Hersch [1955, 1963], Weinberger [1956, 1958, 1974], Hubbard [1961, 1962], Kuttler [1970a,b], and Kreiss [1972].

Remark 10.10. Since the eigenvalue $\lambda_j(h,p)$ are defined by a

Ritz method, they are upper bounds for the exact eigenvalues λ_j:

$$\lambda_j \leq \lambda_j(h,p)$$

(cf. (8.42)). If we could derive a lower bound $\tilde{\lambda}_j(h,p)$, then

one would have bracketed λ_j. Much attention has been directed to

the derivation of lower bounds. Weinstein [1935, 1937, 1953, 1963]

developed the method of intermediate problems. Many authors have

contributed to the development of this and other related varia-

tional methods. We mention D.H. Weinstein [1934], Aronszajn and

Weinstein [1942], Aronszajn [1948, 1949-50], Weinberger [1952a;

1952b, 1956, 1959, 1960], Bazley [1959], Bazley and Fox [1961a,

1961b, 1963]. In addition we mention the monographs by Collatz

[1948], Weinstein & Stenger [1972], and Weinberger [1974].

Remark 10.11. Most books and monographs that treat finite element

methods contain a section or chapter on eigenvalue problems. For

a survey of books and monographs on finite element methods we refer

to Noo. [1985]. Of the more mathematically oriented of these, we

mention Strang and Fix [1973], Oden and Reddy [1976], and Oden and

Carey [1982].

Section 11. Approximation by Mixed Methods

In Section 3 we saw, in terms of an example, how eigenvalue problems can be given mixed formulations. Mixed formulations can be discretized and thereby lead to approximation methods referred to as mixed finite element methods. In this section we discuss three such methods. We begin by presenting an abstract result designed for the analysis of mixed methods.

Remark 11.1. Mixed methods for source problems have received considerable attention. We mention Hermann [1967], Glowinski [1973], Johnson [1973], Oden [1973], Brezzi [1974], Ciarlet and Raviart [1974], Mercier [1974], Scholz [1976], Raviart and Thomas [1977], Brezzi and Raviart [1978], Falk [1978], Babuška, Osborn, and Pitkäranta [1980], and Falk and Osborn [1980].

A. An Abstract Result

Let V, W, H and G be four real Hilbert spaces with inner products and norms $(\cdot, \cdot)_V$, $\|\cdot\|_V$, $(\cdot, \cdot)_W$, $\|\cdot\|_W$, $(\cdot, \cdot)_H$, $\|\cdot\|_H$, and $(\cdot, \cdot)_G$, $\|\cdot\|_G$, respectively. We assume $V \subset H$ and $W \subset G$. Let $A(\sigma, \psi)$ and $B(\psi, u)$ be bilinear forms on $H \times H$ and $V \times W$, respectively, that satisfy

$$(11.1a) \qquad |A(\sigma, \psi)| \le C_1 \|\sigma\|_H \|\psi\|_H, \quad \forall \, \sigma, \psi \in H$$

and

$$(11.1b) \qquad |B(\psi, u)| \le C_2 \|\psi\|_V \|u\|_W, \quad \forall \, \psi \in V, \, u \in W.$$

We assume $A(\sigma, \psi)$ is symmetric and satisfies

$$(11.2a) \qquad A(\sigma, \sigma) > 0, \quad \forall \, 0 \ne \sigma \in H,$$

and assume

172

$(11.2b)$
$$\sup_{\psi \in V} |B(\psi,u)| > 0, \ \forall \ 0 \neq u \in W.$$

We then consider the following eigenvalue problem:

(11.3)
$$\begin{cases} (\sigma,u) \in V \times W, \ (\sigma,u) \neq (0,0) \\ A(\sigma,\psi) + B(\psi,u) = 0, \ \forall \ \psi \in V \\ B(\sigma,v) = -\lambda(u,v)_G, \ \forall \ v \in W \end{cases}$$

A discretization of (11.3) is obtained by selecting finite dimen-sional spaces $V_h \subset V$ and $W_h \subset W$ and considering the approximate eigenvalue problem

(11.4)
$$\begin{cases} (\sigma_h,u_h) \in V_h \times W_h, \ (\sigma_h,u_h) \neq (0,0) \\ A(\sigma_h,\psi) + B(\psi,u_h) = 0, \ \forall \ \psi \in V_h \\ B(\sigma_h,v) = -\lambda_h(u_h,v)_G, \ \forall \ v \in W_h. \end{cases}$$

We then view $(\lambda_h,(\sigma_h,u_h))$ as an approximation to $(\lambda,(\sigma,u))$. Given bases for V_h and W_h, (11.4) becomes a matrix eigenvalue problem.

Remark 11.2. If we let

$$a((\sigma,u),(\psi,v)) = A(\sigma,\psi) + B(\psi,u) + B(\sigma,v),$$
$$b((\sigma,u),(\psi,v)) = -(u,v)_G,$$

and

$$H = V \times W,$$

then (11.3) can be written as

(11.5)
$$\begin{cases} (\sigma,u) \in H, (\sigma,u) \neq (0,0) \\ a((\sigma,u),(\psi,v)) = \lambda b((\sigma,u),(\psi,v)), \ \forall \ (\psi,v) \in H, \end{cases}$$

which has the form of (8.10). Also (11.4) has the form of (8.14)

with $S_{1,h} = S_{2,h} = V_h \times W_h$. (11.3) and (11.5) do not, however, satisfy all of the hypotheses of the results in Section 8. We thus need an alternative analysis. This will be provided by Theorem 11.1, which is based on the results of Section 7. Note that even though the methods considered in this and the next section are not covered by the results of Section 8, it is still useful to discuss them, to the extent possible, in terms of the basic step 1), 1'), 2'), and 3) introduced in Section 10.

In order to estimate the error in the approximate eigenpairs $(\lambda_h, (\sigma_h, u_h))$ we consider the associated source and approximate source problems:

$$(11.6) \quad \begin{cases} \text{Given} \quad g \in G, \quad \text{find} \quad (\sigma, u) \in V \times W \quad \text{satisfying} \\ A(\sigma, \psi) + B(\psi, u) = 0, \quad \forall \; \psi \in V \\ B(\sigma, v) = -(g, v)_G, \quad \forall \; v \in W; \end{cases}$$

$$(11.7) \quad \begin{cases} \text{Given} \quad g \in G, \quad \text{find} \quad (\sigma_h, u_h) \in V_h \times W_h \quad \text{satisfying} \\ A(\sigma_h, \psi) + B(\psi, u_h) = 0, \quad \forall \; \psi \in V_h \\ B(\sigma_h, v) = -(g, v)_G, \quad \forall \; v \in W_h. \end{cases}$$

We assume (11.6) and (11.7) are uniquely solvable for each $g \in G$. We then introduce the corresopnding component solution operators:

$$(11.8a) \quad \begin{cases} S : G \longrightarrow V \\ Sg = \sigma, \end{cases}$$

$$(11.8b) \quad \begin{cases} S_h : G \longrightarrow V \\ S_h g = \sigma_h, \end{cases}$$

$$(11.8c) \quad \begin{cases} T : G \longrightarrow G \\ Tg = u, \end{cases}$$

174

$$(11.8d) \qquad \begin{cases} T_h \; : \; G \longrightarrow G \\ T_h g = u_h, \end{cases}$$

where (σ,u) and (σ_h,u_h) are defined by (11.6) and (11.7), respectively. (Note that the T introduced here is different from that introduced in (8.8).)

The eigenpairs $(\lambda,(\sigma,u))$ of (11.3) can be characterized in terms of the operator T. Before establishing this we note that $\lambda = \dfrac{A(\sigma,\sigma)}{(u,u)_G}$, which shows that $\lambda > 0$. This follows from (11.3) and the observation that both components u and σ of an eigenvector are nonzero. Now, if $(\lambda,(\sigma,u))$ is an eigenpair of (11.3), then $\lambda Tu = u, \; u \neq 0$, and if $\lambda Tu = u, \; u \neq 0$, then there is a $\sigma \in V(\sigma = S(\lambda u))$ such that $(\lambda,(\sigma,u))$ is an eigenpair of (11.3). Thus λ is an eigenvalue of (11.3) if and only if λ^{-1} is an eigenvalue of T. The correspondence between eigenvectors is given by $u \longleftrightarrow (\sigma,u)$. In a similar way the approximate eigenvalues defined by (11.4) can be characterized in terms of the eigenvalues of T_h. λ_h is an eigenvalue of (11.4) if and only if λ_h^{-1} is an eigenvalue of T_h; the correspondence between the eigenpairs is given by $u_h \longleftrightarrow (\sigma_h,u_h)$.

We assume

$$(11.9) \qquad \| T - T_h \|_{GG} \longrightarrow 0 \quad \text{as} \quad h \longrightarrow 0,$$

where, for an operator $A \; : \; D(A)(cX) \longrightarrow Y$, we let

$$\| A \|_{XY} = \sup_{w \in D(A)} \frac{\| Aw \|_Y}{\| w \|_X}.$$

(In particular, we assume T is a bounded operatorn on G.) Since $\dim R(T_h) < \upsilon$ for each h, the T_h are compact and (11.9) thus

175

implies T is compact. We also note that T is selfadjoint on G. This is seen as follows. Let $v = Tf$ in the second equation in (11.6) to obtain

$$B(Sg, Tf) = -(g, Tf)_G.$$

Again consider (11.6), but with g replaced by f, and let $v = Sg$ in the first equation to get

$$A(Sf, Sg) + B(Sg, Tf) = 0.$$

From these two equations we have

(11.10) $$(g, Tf)_G = A(Sf, Sg), \ \forall \ f, g \in G.$$

Using (11.10) and the symmetry of A we get

$$(Tg, f)_G = (f, Tg)_G = A(Sg, Sf) = A(Sf, Sg) = (g, Tf)_G,$$

showing T is selfajoint. In a similar way we see T_h is self-adjoint.

We now apply Theorems 7.3 and 7.4 to the operator T and family of operators $\{T_h\}$ on the space G. By virtue of the correspondence between the eigenpairs of T and T_h and those of (11.3) and (11.4) we will thereby obtain estimates for the errors in $(\lambda_h, (\sigma_h, u_h))$. The hypotheses have all been shown to be satisfied; cf. Remarks 7.5 and 7.6. Let λ^{-1} be an eigenvalue of multiplicity m. Since $\|T - T_h\|_{GG} \rightarrow 0$ we know that m eigenvalues $\lambda_1(h)^{-1}, \ldots, \lambda_m(h)^{-1}$ of T_h converge to λ^{-1}. Since T and T_h are selfadjoint the relevant ascents are one and all eigenvalues have equal geometric and algebraic multiplicities. Let $\bar{M}(\lambda^{-1})$ be the eigenspace of T corresponding to λ^{-1}. Recall that $\bar{M} =$

$\bar{M}(\lambda^{-1}) = R(E)$, the range of the spectral projection E associated with T and λ^{-1}. We have denoted this space by $\bar{M}$ to distinguish it from the set M of normalized eigenvectors introduced in Section 8.

Theorem 11.1. Under the assumptions made above, there is a constant C such that

$$|\lambda - \lambda_\ell(h)| \leq C\{\|(S-S_h)|_{\bar{M}}\|^2_{GH} + \|(S-S_h)|_{\bar{M}}\|_{GV}\|(T-T_h)|_{\bar{M}}\|_{GW}$$
(11.11)
$$+ \|(T-T_h)|_{\bar{M}}\|^2_{GG}\}, \quad \ell = 1, \ldots, m.$$

Proof. Let $u_1, \ldots, u_m$ be an orthonormal basis for $\bar{M}(\lambda^{-1})$. From Theorem 7.3 with $\alpha = 1$ we have

$$|\lambda^{-1} - \lambda_\ell(h)^{-1}| \leq C\{\sum_{i,j=1}^{m} |((T-T_h)u_i, u_j)_G| + \|(T-T_h)|_{\bar{M}}\|^2_{GG}\},$$
(11.12)
$$\ell = 1, \ldots, m.$$

For $g, f \in G$ we estimate $((T - T_h)g, f)_G$. Adding the two equations in (11.6) and recalling the definitions of Tg and Sg in (11.8) we find

$$(g,v)_G = -A(Sg,\psi) - B(\psi,Tg) - B(Sg,v), \quad \forall \ (\psi,v) \in V \times W.$$

Setting $v = (T-T_h)f$ and $w = (S-S_h)f$ yields

(11.13) $\quad (g,(T-T_h)f)_G = -A(Sg,(S-S_h)f) - B((S-S_h)f,Tg)$
$$- B(Sg,(T-T_h)f).$$

Next note that substraction of the equations (11.7) from (11.6) with g replaced by f) gives

$$(11.14) \quad A((S-S_h)f,\psi) + B(\psi,(T-T_h)f) + B((S-S_h)f,v) = 0,$$

$$(\psi,v) \in V_h \times W_h.$$

Now, combining (11.13) and (11.14) and using (11.1) we have

$$|(g,(T-T_h)f)_G| = |A((S-S_h)f,Sg-\psi) + B((S-S_h)f,Tg-v)$$

$$+ B(Sg-\psi,(T-T_h)f)|$$

$$\leq C_1\|(S-S_h)f\|_H\|Sg-\psi\|_H$$

$$+ C_2\|(S-S_h)f\|_V\|Tg-v\|_W$$

$$+ C_2\|Sg-\psi\|_V\|(T-T_h)f\|_W.$$

Setting $\psi = S_h g$ and $v = T_h g$ gives

$$|((T-T_h)g,f)_G| \leq C_1\|(S-S_h)f\|_H\|(S-S_h)g\|_H$$

$$(11.15) \qquad\qquad + C_2\|(S-S_h)f\|_V\|(T-T_h)g\|_W$$

$$+ C_2\|(S-S_h)g\|_V\|(T-T_h)f\|_W.$$

Letting $g = u_i$ and $f = u_j$ in (11.15) yields

$$(11.16) \qquad |((T-T_h)u_i,u_j)_G| \leq C_1\|(S-S_h)\|_{\overline{M}}^2 {}_{GH}$$

$$+ 2C_2\|(S-S_h)\|_{\overline{M}} {}_{GV}\|(T-T_h)\|_{\overline{M}} {}_{GW}.$$

(11.11) follows immediately from (11.12) and (11.16). $\qquad\qquad \square$

Theorem 11.2. Under the assumptions made above, there is a constant C such that

$$(11.17) \qquad\qquad \|u-u_h\|_G \leq C\|(T-T_h)\|_{\overline{M}} {}_{GG}.$$

Proof. This result is an immediate consequence of Theorems 7.1

and 7.4. Note that we have given this estimate the simplified

form it has when λ is simple, and it would have to be modified

in the general case. Cf. the statement of Theorem 10.1 and (8.44)

– (8.46).   □

Theorems 11.1 and 11.2 were proved by Osborn [1979] and by

Mercier, Osborn, Rappaz, and Raviart [1981].

B. A Mixed Method for the Vibrating Membrane

We consider, as in Subsection 10.B., the vibrating mem-

brane problem

(11.18)
$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega \\ u = 0, & \text{on } \Gamma = \partial\Omega, \end{cases}$$

where Ω is a convex polygon in R^2, but we will here give it a

mixed variational formulation. Otherwise we will proceed in a

parallel way, discussing in turn the steps 1), 1'), 2), and 3)

introducted in Subsection 10.A. We will clearly see how the vari-

ational formulation influences the entire approximation method.

Before proceeding with the variational formulation, we intro-

duce an additional function space. Let

$$H(\text{div},\Omega) = \{\sigma = (\sigma_1,\sigma_2) : \sigma_1,\sigma_2 \in H^0(\Omega) \text{ and there exists}$$
$$z = \text{div } \sigma \in H^0(\Omega) \text{ such that } \int_\Omega \sigma\cdot\nabla\phi \ dxdy =$$
$$-\int_\Omega z\phi \ dxdy, \ \forall \ \phi \ \in \ C_0^\infty(\Omega)\},$$

$$\|\sigma\|^2_{H(\text{div},\Omega)} = \int_\Omega [\sigma_1^2 + \sigma_2^2 + (\text{div } \sigma)^2]dxdy.$$

1) Variational Formulation

Suppose (λ,u) is an eigenpair of (11.18), by which we will

mean

$$(11.19) \quad \begin{cases} 0 \neq u \in H_0^1(\Omega) \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \ dxdy = \lambda \int_\Omega uv \ dxdy, \ \forall \ v \in H_0^1(\Omega), \end{cases}$$

i.e., we will assume (11.18) to have the variational formulation considered in Subsection 10.B. We now derive a mixed variational formulation for (11.18). Introduce the auxiliary variable

$$(11.20) \qquad\qquad\qquad \sigma = \nabla u.$$

From (11.19) we see that $\sigma \in H(\text{div},\Omega)$ and

$$(11.21) \qquad\qquad\qquad \text{div} \ \sigma = -\lambda u.$$

From (11.21) we get

$$(11.22) \qquad \int_\Omega v \ \text{div} \ \sigma \ dxdy = -\lambda \int_\Omega uv \ dxdy, \ \forall \ v \in H^0(\Omega)$$

and from (11.20) and the definition of $H(\text{div},\Omega)$ we have

$$(11.23) \qquad \int_\Omega \sigma \cdot \psi \ dxdy = \int \nabla u \cdot \psi \ dxdy = -\int_\Omega u \ \text{div} \ \psi \ dxdy,$$

$$\forall \ \psi \in H(\text{div},\Omega).$$

Combining (11.22) and (11.23) we obtain

$$(11.24) \quad \begin{cases} (\sigma,u) \in H(\text{div},\Omega) \times H^0(\Omega), \ (\sigma,u) \neq (0,0) \\ \displaystyle\int_\Omega \sigma \cdot \psi \ dxdy + \int_\Omega u \ \text{div} \ \psi \ dxdy = 0, \ \forall \ \psi \in H(\text{div},\Omega) \\ \displaystyle\int_\Omega v \ \text{div} \ \sigma \ dxdy = -\lambda \int_\Omega uv \ dxdy, \ \forall \ v \in H^0(\Omega). \end{cases}$$

Now suppose $(\lambda,(\sigma,u))$ satisfies (11.24). Let $\bar{u}$ be the

180

solution to

$$(11.25) \qquad \begin{cases} \Delta \bar{u} = \lambda u \quad \text{in} \quad \Omega \\ \bar{u} = 0 \quad \text{in} \quad \Gamma \end{cases}$$

and let

$$\bar{\sigma} = \nabla \bar{u}.$$

Then, by the argument used above,

$$(11.26) \qquad \begin{cases} (\bar{\sigma}, \bar{u}) \in H(div, \Omega) \times H^0(\Omega) \\ \int_{\Omega} \bar{\sigma} \cdot \psi \; dxdy + \int_{\Omega} \bar{u} \; div \; \psi \; dxdy = 0, \; \forall \; \psi \in H(div, \Omega) \\ \int_{\Omega} v \; di\bar{v} \; \sigma \; dxdy = -\lambda \int_{\Omega} uv \; dxdy, \; \forall \; v \in {}^0 H (\Omega). \end{cases}$$

Subtraction of the equations in (11.26) from those in (11.24) yields

$$(11.27) \qquad \begin{cases} (\sigma - \bar{\sigma}, \; u - \bar{u}) \in H(div, \Omega) \times H^0(\Omega) \\ \int_{\Omega} (\sigma - \bar{\sigma}) \cdot \psi \; dxdy + \int_{\Omega} (u - \bar{u}) div \; \psi \; dxdy = 0, \; \forall \; \psi \in H(div, \Omega) \\ \int_{\Omega} div(\sigma - \bar{\sigma}) v \; dxdy = 0, \; \forall \; v \in H^0(\Omega). \end{cases}$$

In (11.27), if in the second equation we take v arbitrary in $H^0(\Omega)$ we get $div(\sigma - \bar{\sigma}) = 0$, and if we then take $\psi = \sigma - \bar{\sigma}$ in the first equation we obtain $0 = \int_{\Omega} (\sigma - \bar{\sigma}) \cdot (\sigma - \bar{\sigma}) dxdy$, which implies

$$(11.28) \qquad \sigma = \bar{\sigma}.$$

Then the first equation in (11.27) implies

$$(11.29) \qquad \int_{\Omega} (u - \bar{u}) \; div \; \psi \; dxdy = 0, \; \forall \; \psi \in H(div, \Omega).$$

Let w satisfy $\Delta w = u - \bar{u}$ and let $\psi = \nabla w$ in (11.29). Since div $\psi = u - \bar{u}$, this choice leads to

$$(11.30) \qquad\qquad\qquad u = \bar{u}$$

(11.25), (11.28), and (11.30) show that (λ, u) is an eigenpair of (11.18) (or (11.19)), and that $\sigma = \nabla u$.

In summary, if (λ, u) is an eigenpair of (11.18) and $\sigma = \nabla u$, then $(\lambda, (u, \sigma))$ satisfies (11.24), and if $(\lambda, (\sigma, u))$ satisfies (11.24), then (λ, u) is an eigenpair of (11.18) and $\sigma = \nabla u$. (11.24) is the desired mixed formulation.

It is immediate that (11.24) has the form of (11.3) with

$$V = H(\text{div}, \Omega),$$

$$W = G = H^0(\Omega),$$

$$H = H^0(\Omega),$$

$$A(\sigma, \psi) = \int_\Omega \sigma \cdot \psi \; dxdy,$$

and

$$B(\psi, u) = \int_\Omega u \; \text{div} \; \psi \; dxdy.$$

Furthermore, A is symmetric and (11.1) and (11.2) hold. The symmetry of A and (11.1) and (11.2a) are trivial. To prove (11.2b), let w solve $\Delta w = u$ and set $\tilde{\psi} = \nabla w$. Then div $\tilde{\psi} = u$ and we have

$$\sup_{\psi \in H(\text{div}, \Omega)} |\int_\Omega u \; \text{div} \; \psi \; dxdy| \geq |\int_\Omega u \; \text{div} \; \tilde{\psi} \; dxdy|$$

$$= \int_\Omega u^2 \; dxdy$$

182

$$> 0, \quad \text{for} \quad 0 \ne u \in H^0(\Omega),$$

which proves (11.2b).

From the fact that (11.18) has a sequence of positive eigen-values and from the correspondence between the eigenpairs of (11.18) and (11.24) we see that (11.24) has a sequence of eigen-values

$$0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow \infty$$

and corresponding eigenfunctions

$$(\sigma_1, u_1), \ (\sigma_2, u_2), \ldots,$$

with $\sigma_j = \nabla u_j$ and with the (λ_j, u_j) being the eigenpairs of (11.18).

1′) <u>Regularity</u> <u>of</u> <u>the</u> <u>Eigenfunctions</u>

If (σ, u) is an eigenfunction of (11.24), then u is an eigenfunction of (11.18) and $\sigma = \nabla u$. Thus the regularity of (σ, u) can be inferred from the regularity of the eigenfunction of (11.16), which was discussed in Subsection 11.B.

2) <u>Discretization</u> <u>of</u> (11.24) <u>and</u> <u>Assessment</u> <u>of</u> <u>the</u> <u>Accuracy</u> <u>of</u> <u>the</u> <u>Approximate</u> <u>Eigenvalues</u> <u>and</u> <u>Eigenfunctions</u>

We will use a discretization of the general form of (11.4). It thus remains to select the subspaces $V_h \subset H(\text{div}, \Omega)$ and $W_h \subset H^0(\Omega)$. This will be done with an eye toward ensuring (11.9) holds and the terms on the right side of (11.11) in Theorem 11.1 are small. A mixed method approximation of the associated source problem (cf. (11.6) and (11.7)) has been proposed and analyzed in Raviart and Thomas [1977]. We will take their choice of trial and

test functions. The source problem has also been analyzed by Falk and Osborn [1980].

Let $\hat{T}$ be the unit triangle in the (ξ, η)-plane whose vertices are $\hat{a}_1 = (1,0)$, $\hat{a}_2 = (0,1)$, and $\hat{a}_3 = (0,0)$. Then with $p \geq 0$ an even integer and $\hat{T}$ associate the space $\hat{Q}^{(p+1)}$ of all functions $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2)$ of the form

$$\hat{\psi}_1 = \text{pol}_p(\xi, \eta) + \alpha_0 \xi^{p+1} + \alpha_1 \xi^p \eta + \ldots + \alpha_{p/2} \xi^{p/2+1} \eta^{p/2},$$

$$\hat{\psi}_2 = \text{pol}_p(\xi, \eta) + \beta_0 \eta^{p+1} + \beta_1 \xi \eta^p + \ldots + \beta_{p/2} \xi^{p/2} \eta^{p/2+1},$$

where $\text{pol}_p(\xi, \eta)$ denotes an arbitrary polynomial of degree p and where

$$\sum_{i=0}^{p/2} (-1)^i (\alpha_i - \beta_i) = 0,$$

and with $p \geq 1$ an odd integer and $\hat{T}$ associate the space $\hat{Q}^{(p+1)}$ of all $\hat{\psi}$ of the form

$$\hat{\psi}_1 = \text{pol}_p(\xi, \eta) + \alpha_0 \xi^{p+1} + \alpha_1 \xi^p \eta + \ldots + \alpha_{(p+1)/2} \xi^{(p+1)/2} \eta^{(p+1)/2},$$

$$\hat{\psi}_2 = \text{pol}_p(\xi, \eta) + \beta_0 \eta^{p+1} + \beta_1 \xi \eta^p + \ldots + \beta_{(p+1)/2} \xi^{(p+1)/2} \eta^{(p+1)/2},$$

where

$$\sum_{i=0}^{(p+1)/2} (-1)^i \alpha_i = \sum_{i=0}^{(p+1)/2} (-1)^i \beta_i = 0.$$

We remark that for $\hat{\psi} \in \hat{Q}^{(p+1)}$, $\hat{\psi}_1$ and $\hat{\psi}_2$ are polynomials of degree $p+1$. With a general triangle T in the (x,y)-plane, we associate the space $Q_T^{(p+1)}$ defined by

$$Q_T^{(p+1)} = \{\psi \; : \; \psi(x,y) = \frac{1}{J_T} B_T \hat{\psi}(F_T^{-1}(x,y)), \; \hat{\psi} \in Q^{(p+1)}\},$$

where $F_T(\xi,\eta) = B_T(\xi,\eta) + b_T$ is the linear transformation mapping $\hat{T}$ onto T and $J_T = \det(B_T)$.

Let $\gamma = \{\tau\} = \{\tau_h\}$ be a family of α-regular triangularization of $\bar{\Omega}$. Then for $p \geq 0$ an integer let

(11.31a) $\quad V_h = \{\psi \in H(\mathrm{div},\Omega) \; : \; \psi|_T \in Q_T^{(p+1)}, \; \forall \; T \in \tau_h\}$

and

(11.31b) $W_h = \{u \in H^0(\Omega) \; : \; u|_T$ polynomial of degree $p, \; \forall \; T \in \tau_h\}$.

Now we consider (11.4) with this choice for V_h and W_h. (11.4) will have eigenvalues

$$\lambda_{1,h} \leq \cdots \leq \lambda_{N,h}$$

and corresponding eigenfunctions

$$(\sigma_{1,h}, u_{1,h}), \ldots, (\sigma_{N,h}, u_{N,h}),$$

where $N = \dim(V_h \times W_h)$. It remains to derive error estimates by applying Theorems 11.1 and 11.2.

<u>Theorem 11.3</u>. Let V_h and W_h be selected as in (11.31). Suppose the eigenfunctions of (11.18) belong to $H^{p+2}(\Omega)$. Then

(11.32) $$|\lambda_{k,h} - \lambda_k| \leq C(p)h^{2p+2}$$

and

(11.33) $$\|u_{k,h} - u_k\|_{0,\Omega} \leq C(p)h^{p+1}.$$

<u>Proof</u>. We begin by showing that all of the hypotheses of Theorems 11.1 and 11.2 are satisfied. We have already noted that A is

symmetric and that (11.1) and (11.2) are satisfied for the problem (11.24).

The source problem (11.6) is uniquely solvable for each $g \in$ $G = H^0(\Omega)$. In fact the unique solution is (σ, u), where

$$\begin{cases} -\Delta u = g \\ u \in H_0^1(\Omega) \end{cases}$$

and

$$\sigma = \nabla u$$

(cf. the discussion in 1) above). To see that (11.7) is uniquely solvable for each $g \in G$ it is sufficient to show that $g = 0$ implies σ_h and u_h are zero. Now $g = 0$ implies, using the second equation in (11.7), that $B(\sigma_h, v) = 0$, $\forall \, v \in W_h$. Setting $\psi = \sigma_h$ in the first equation and using this fact shows that $A(\sigma_h, \sigma_h) = 0$ which, together with (11.2a), shows that $\sigma_h = 0$. Then, using the first equation in (11.6) again we get $B(\psi, u_h) = 0$, $\forall \, \psi \in V_h$. For our specific problem this is $\int_\Omega u_h \, \text{div} \, \psi \, dxdy = 0$, $\forall \, \psi \in V_h$. It is shown in Raviart and Thomas [1977, Theorem 4] that corresponding to any $u_h \in W_h$ there is a $\psi \in V_h$ such that $\text{div} \, \psi_h = u_h$. Using this ψ we thus have $\int_\Omega |u_h|^2 dxdy = 0$ which implies $u_h = 0$.

It remains to check (11.9). Falk and Osborn [1980, Section 3(d)] have shown that

(11.34)
$$\|Tg - T_h g\|_{0,\Omega} \leq \begin{cases} Ch^2 \|Tg\|_{2,\Omega}, & \text{for} \quad p \geq 1 \\ Ch \, \|Tg\|_{2,\Omega}, & \text{for} \quad p = 0 \end{cases}$$

$$\leq C \, h \, |g|_{0,\Omega}, \quad \text{for} \quad p \geq 0,$$

which proves (11.9).

We now apply Theorems 11.1 and 11.2. From Raviart and Thomas [1977, Theorem 5] we have

$$\|(S-S_h)g\|_{H^0(\Omega)} \leq \|(S-S_h)g\|_{H(div,\Omega)} \leq Ch^{p+1}(\|Tg\|_{p+2,\Omega} + \|g\|_{p+1,\Omega})$$

and

$$\|(T-T_h)g\|_{0,\Omega} \leq Ch^{p+1}(\|Tg\|_{p+2,\Omega} + \|g\|_{p+1,\Omega}).$$

If $g \in \bar{M}(\lambda_k^{-1})$, then $Tg = \lambda_k^{-1}g$ and g is an eigenfunction of (11.18) corresponding to λ_k and by our hypotheses, $\|g\|_{p+2,\Omega} < \infty$. Thus

$$(11.35a) \qquad \|(S-S_h)\big|_{\bar{M}}\|_{H^0(\Omega),H^0(\Omega)} \leq Ch^{p+1},$$

$$(11.35b) \qquad \|(S-S_h)\big|_{\bar{M}}\|_{H^0(\Omega),H(div,\Omega)} \leq Ch^{p+1},$$

and

$$(11.35c) \qquad \|(T-T_h)\big|_{\bar{M}}\|_{H^0(\Omega),H^0(\Omega)} \leq Ch^{p+1}.$$

(11.32) follows immediately from Theorem 11.1 and estimates (11.35). (11.33) follows immediately from Theorem 11.2 and (11.35c).

Remark 11.3. Theorems 11.1 and 11.2 estimate the errors in mixed method approximation of eigenpairs in terms of error estimates for the corresponding source problems. For our problem, these were mainly provided by the results of Raviart and Thomas [1977]. Note, however, that estimate (11.34) — the estimate that ensures the approximate eigenvalues converge — is not proved in Raviart and Thomas [1977].

3) Solution of Matrix Eigenvalue Problem

The matrix problem (11.4) with V_h and W_h given in (11.31) is large and sparse, but is not positive definite.

C. A Mixed Method for the Vibrating Plate

The eigenvalue problem

$$(11.36) \qquad \begin{cases} \Delta^2 u = \lambda u, & \text{in } \Omega \\ u = \dfrac{\partial u}{\partial n} = 0, & \text{in } \Omega \end{cases}$$

arises in connection with the small, transverse vibration of a clamped plate. A commonly used variational formulation of (11.36) is

$$(11.37) \qquad \begin{cases} u \in H_0^2(\Omega), \ u \neq 0 \\ \displaystyle\int_\Omega \Delta u \, \Delta v \ dxdy = \lambda \int_\Omega uv \ dxdy, \ \forall \ v \in H_0^2(\Omega). \end{cases}$$

A finite element method based on (11.37) would require trial and test space that were subspaces of $H_0^2(\Omega)$, and this would require c^1-elements, i.e., piecewise polynomials that are c^1 across inter-element boundaries. In order to avoid this requirement we will use a different variational formulation for (11.37), one that permits the use of c^0-elements. We do, however, use (11.37) to show that (11.36) has a sequence of eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow \infty$$

and corresponding eigenfunctions

$$u_1, u_2, \ldots,$$

which can be chosen so that

$$\int_{\Omega} \Delta u_i \, \Delta u_j \, dxdy = \lambda_j \int_{\Omega} u_i u_j \, dxdy = \delta_{ij}.$$

1) Variational Formulation

Introduce the auxiliary variable $\sigma = -\Delta u$. Then (11.36) can be written as a second order system:

$$\begin{cases} \sigma + \Delta u = 0, & \text{in } \Omega \\ -\Delta \sigma = \lambda u, & \text{in } \Omega \\ u = \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma. \end{cases}$$

Multiplying the first equation by ψ, the second by v, integrating over Ω, and integrating by parts leads to

$$0 = \int_{\Omega} \sigma\psi \, dxdy + \int_{\Omega} \Delta u \psi \, dxdy$$

$$= \int_{\Omega} \sigma\psi \, dxdy - \int_{\Omega} \nabla u \cdot \nabla \psi + \int_{\Gamma} \frac{\partial u}{\partial n} \, \psi \, ds$$

$$= \int_{\Omega} \sigma\psi \, dxdy - \int_{\Omega} \nabla u \cdot \nabla \psi \, dxdy, \ \forall \ \psi \in H^1(\Omega)$$

and

$$\lambda \int_{\Omega} uv \, dxdy = - \int \Delta\sigma v \, dxdy$$

$$= \int_{\Omega} \nabla\sigma \cdot \nabla v \, dxdy - \int_{\Gamma} \frac{\partial\sigma}{\partial n} \, v \, ds$$

$$= \int_{\Omega} \nabla\sigma \cdot \nabla v \, dxdy, \ \forall \ v \in H_0^1(\Omega).$$

Thus we arrive at the variational formulation

$$(11.38) \quad \begin{cases} (\sigma, u) \in H^1(\Omega) \times H_0^1(\Omega), \ (\sigma, u) \neq (0,0) \\ \displaystyle\int_\Omega \sigma \psi \ dxdy - \int_\Omega \nabla u \cdot \nabla \psi \ dxdy = 0, \ \forall \ \psi \in H^1(\Omega) \\ \qquad\qquad - \int \nabla \sigma \cdot \nabla v \ dxdy = -\lambda \int_\Omega uv \ dxdy, \ \forall \ v \in H_0^1(\Omega). \end{cases}$$

We derived (11.38) formally from (11.36). One can, however, easily make the argument rigorous with the aid of a well-known regularity result: If w is the solution to

$$\begin{cases} \Delta^2 w = f, \quad \text{in} \ \Omega \\ w = \dfrac{\partial w}{\partial n} = 0, \quad \text{on} \ \Gamma, \end{cases}$$

where Ω is a convex polygon and $f \in H^0(\Omega)$, then $w \in H^3(\Omega)$ and $\|w\|_{3,\Omega} < C\|f\|_{0,\Omega}$, cf. Grisvard [1985] and Kellogg and Osborn [1975]. We assume Ω is a convex polygon in the remainder of this subsection. Using this result we can show that if (λ, u) is an eigenpair of (11.36) and $\sigma = -\Delta u$, then $(\lambda, (\sigma, u))$ is an eigenpair of (11.38), and if $(\lambda, (\sigma, u))$ is an eigenpair of (11.38), then (λ, u) is an eigenpair of (11.36) and $\sigma = -\Delta u$. (11.38) has the form of (11.3) with

$$V \quad H^1 \ldots$$

$$W \quad H_0^1 \ldots$$

$$H \quad \ldots = H \ldots$$

$$A \ldots = \ldots dxdy.$$
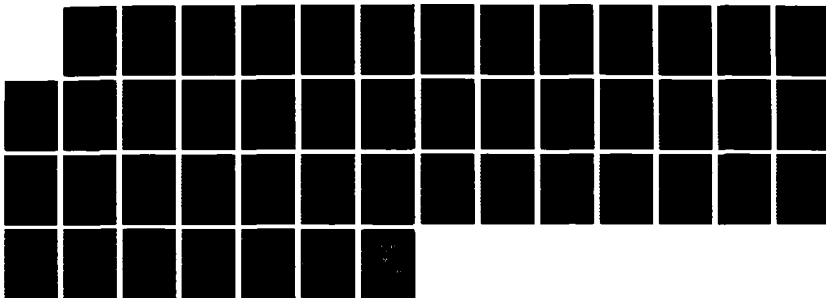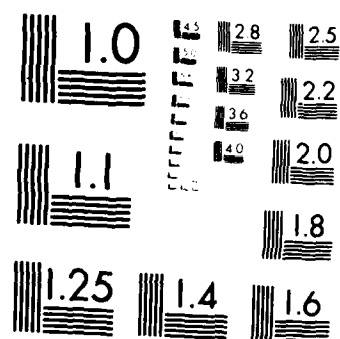
MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

$$B(\psi,u) = \int_{\Omega} \nabla\psi\cdot\nabla u \; dxdy.$$

It is easily seen that A is symmetric and that (11.1) and (11.2) are satisfied.

(11.38) has eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow \infty$$

and corresponding eigenfunctions

$$(\sigma_1,u_1), \; (\sigma_2,u_2),\ldots,$$

with $\sigma_j = -\Delta u_j$.

1') Regularity of the Eigenfunctions

If (σ,u) is an eigenfunction of (11.38) then, as we have seen above, u is an eigenfunction of (11.36) and $\sigma = -\Delta u$, and hence the regularity of (σ,u) can be inferred from the regularity properties of (11.36). For results on this later regularity question we refer to Grisvard [1985] and Kellogg and Osborn [1975].

2) Discretization of (11.38) and Assessment of the Accuracy of the Approximate Eigenpairs

As in Subsection B. above, our discretization will be via (11.4). For our specific problem, a mixed method for the associated source problem has been studied by Glowinski [1973], Ciarlet and Raviart [1974], Mercier [1974], and Falk and Osborn [1980]. We will use the same trial and test spaces employed in those papers.

Let $\gamma = \{\tau\} = \{\tau_h\}$ be a family of α-regular, q-quasiuniform triangulations of $\bar{Q}$. Then for $p = 2,3,\ldots,$ let

(11.39a) $$V_h = S^p(\tau_h)$$

and

(11.39b) $$W_h = S_0^p(\tau_h) \cap H_0^1(\Omega).$$

We then consider (11.4) with these choices. We will have approximate eigenvalues and eigenfunctions

$$\lambda_{1,h} \leq \ldots \leq \lambda_{N,h}$$

and

$$(\sigma_{1,h}, u_{1,h}), \ldots, (\sigma_{N,h}, u_{N,h}),$$

where $N = \dim(V_h \times W_h)$.

<u>Theorem 11.4</u>. Let V_h and W_h be as in (11.39) with $p \geq 2$ and suppose the eigenfunctions of (11.36) belong to $H^{p+1}(\Omega)$. Then

(11.40) $$|\lambda_{k,h} - \lambda_k| \leq C(p)h^{2p-2}$$

and

(11.41) $$\|u_{k,h} - u_k\|_{0,\Omega} \leq C(p)h^p.$$

<u>Proof</u>. The symmetry of A and the validity of (11.1) and (11.2) for problem (11.38) have already been noted.

The source problem (11.6) is uniquely solvable for each $g \in G = H^0(\Omega)$. The unique solution is (σ, u), where

$$\begin{cases} \Delta^2 u = g \\ u \in H_0^2(\Omega) \end{cases}$$

and

$$\sigma = -\Delta u$$

(cf. the derivation of (11.38)). The unique solvability of (11.7) is easily checked.

192

Falk and Osborn [1980, Section 3a] have shown that

$$\|Tg - T_hg\|_{0,\Omega} \leq Ch^2\|Tg\|_{3,\Omega}.$$

This, together with the regularity result mentioned above, gives

$$\|(T - T_h)g\|_{0,\Omega} \leq Ch^2\|g\|_{0,\Omega},$$

which proves (11.9)

Thus, all of the hypotheses for Theorems 11.1 and 11.2 have been verified for the problem under consideration. Using the results in Falk and Osborn [1980, Section 3a], we have

$$\|(S-S_h)g\|_{0,\Omega} \leq Ch^{p-1}\|Tg\|_{p+1,\Omega},$$

$$\|(S-S_h)g\|_{1,\Omega} \leq Ch^{p-2}\|Tg\|_{p+1,\Omega},$$

$$\|(T-T_h)g\|_{0,\Omega} \leq Ch^{p}\|Tg\|_{p+1,\Omega},$$

and

$$\|(T-T_h)g\|_{1,\Omega} \leq Ch^{p}\|Tg\|_{p+1,\Omega},$$

from which we obtain

(11.42a) $$\|(S-S_h)|_{\bar{M}}\|_{H^0(\Omega),H^0(\Omega)} \leq Ch^{p-1},$$

(11.42b) $$\|(S-S_h)|_{\bar{M}}\|_{H^0(\Omega),H^1(\Omega)} \leq Ch^{p-2},$$

(11.42c) $$\|(T-T_h)|_{\bar{M}}\|_{H^0(\Omega),H^0(\Omega)} \leq Ch^{p},$$

and

(11.42d) $$\|(T-T_h)|_{\bar{M}}\|_{H^0(\Omega),H^1_0(\Omega)} \leq Ch^{p}.$$

(11.40) follows immediately from Theorem 11.1 and (11.42), and (11.41) follows from Theorem 11.2 and (11.42c).

Remark 11.4. The estimates obtained in this subsection were first obtained by Canuto [1978]. We note, however, that the estimation techniques used here will yield an improvement over the estimates of Canuto in the case when the eigenfunctions have low regularity. Our method of proof does not yield any estimates for $p = 1$. For this case, see Ishihara [1978 a,b].

3) Solution of Matrix Eigenvalue Problem

See subsection A.3) above.

For further results in eigenvalue approximation by mixed methods, and also by hybrid methods, we refer to Mercier, Osborn, Rappaz, and Raviart [1981], Mercier and Rappaz [1978], and Ishihara [1977].

Remark 11.5. We have seen in this Section and in Section 11 that there are various methods available for the approximate calculation of the eigenvalues of a specific problem. For example, we have analyzed two methods for the membrane problem. Furthermore, this discussion, together with that in Section 3, shows that there are many more possibilities. Clearly the rational choice of a method for any particular concrete problem is important. The effective choice of a method is complex, depending on many aspects of the underlying problem.

D. A Mixed Method for a Problem in One Dimension with Rough Coefficients

Consider the problem

$$(11.43) \qquad \begin{cases} -(a(x)u')' = \lambda b(x)u, \ 0 < x < 1 \\ u(0) = u(1) = 0. \end{cases}$$

This is a special case of the problem (1.8), (1.9a) discussed in Section 1. We will be especially interested here in the case in which the coefficients $a(x)$ and $b(x)$ are rough functions. Such problems arise in the analysis of the vibrations of structures with rapidly varying material properties, of composite materials, for example. In Section 3, we gave (11.43) the mixed formulation (3.26) (or (3.27)). In this subsection we analyze a mixed method based on (3.27).

Hence we consider the problem

$$(11.44) \quad \begin{cases} (\sigma, u) \in L_2(0,1) \times H_0^1(0,1), (\sigma, u) \neq (0,0) \\[2ex] \displaystyle\int_0^1 \frac{\sigma \psi}{a} \, dx - \int_0^1 u' \psi \, dx = 0, \ \forall \ \psi \in L_2(0,1) \\[2ex] \displaystyle- \int_0^1 \sigma v' \, dx = -\lambda \int_0^1 buv \, dx, \ \forall \ v \in H_0^1(0,1), \end{cases}$$

where $a(x)$ is of bounded variation and $b(x)$ is measurable and

$$0 < a_0 \leq a(x) \leq a_1, \ 0 < b_0 \leq b(x) \leq b_1.$$

(11.44) is of the form (11.3) with

$$V = H = L_2(0,1),$$

$$W = H_0^1(0,1),$$

$$G = L_2(0,1), \quad \text{with} \quad (u,v)_G = \int_0^1 buv \, dx,$$

$$A(\sigma, \psi) = \int_0^1 \frac{\sigma \psi}{a} \, dx,$$

and

195

$$B(\sigma,v) = -\int_0^1 \sigma v' \, dx.$$

$(\lambda,(\sigma,u))$ is an eigenpair of (11.44) if and only if (λ,u) is an eigenpair of (11.43) and $\sigma = au'$. We discretize (11.44) by letting $\tau = \{T_i\}_{i=1}^{M(\tau)}$ be a mesh on $[0,1]$, defining

(11.45a) $\quad V_h = \{\sigma : \sigma|_{T_i} = $ a constant, $i = 1,\ldots,M(\tau)\}$

and

(11.45b) $W_h = \{v : v \in H_0^1(0,1), v|_{T_i} = $ a linear polynomial,

$$i = 1,\ldots,M(\tau)\},$$

with $h = h(\tau) = \max\limits_{i=1,\ldots,M(\tau)}$ diam T_i, and considering (11.4).

The eigenpairs $(\lambda_h,(\sigma_h,u_h))$ of (11.4) are then considered as approximations to the eigenpairs $(\lambda,(\sigma,u))$ of (11.44). Although this approximation method satisfies the hypotheses of Theorem 11.1, a direct application of that result does not yield the best possible estimate. We will employ an analysis that is parallel to, but different than, that used in the proof of Theorem 11.1.

The analysis begins by introducing the operators $T,T_h : G \rightarrow G$ and $S,S_h : G \rightarrow V$ that are defined in (11.8). λ is an eigenvalue of (11.44) if and only if λ^{-1} is an eigenvalue of T; the correspondence between the eigenfunctions is given by $(\sigma,u) \leftrightarrow u$. Likewise λ_h is an eigenvalue of (11.4) if and only if λ_h^{-1} is an eigenvalue of T_h, with the correspondence between eigenfunctions given by $(\sigma_h,u_h) \leftrightarrow u_h$. $\|T-T_h\|_{GG} \rightarrow 0$, as will be shown later, so we may apply Theorem 7.2 to T and T_h on the space G. Let λ^{-1} be an eigenvalue of T. The eigenvalues of a problem of the type (11.43) are simple and hence λ^{-1} is a simple

196

eigenvalue of T. Thus one eigenvalue λ_h of (11.4) converges to λ. By Theorem 7.2 we have

$$(11.46) \qquad |\lambda - \lambda_h| \leq C\{|((T-T_h)u,u)_G| + \|(T-T_h)u\|_G^2\},$$

where u is any eigenfunction of T corresponding to λ^{-1} with $\|u\|_G = 1$. We now proceed to analyze $((T-T_h),u)_G$.

From (11.6) we have

$$((T-T_h)u,u)_G = \int_0^1 bu(T-T_h)u \; dx$$

$$(11.47) \qquad\qquad = -B(Su,(T-T_h)u)$$

$$\qquad\qquad = A(Su,(S-S_h)u) + B(((S-S_h)u,Tu) - B(Su,(T-T_h)u),$$

and from (11.6) and (11.7) we get

$$(11.48) \qquad 0 = A((S-S_h)u,\xi) + B((S-S_h)u,\eta) + B(\xi,(T-T_h)u),$$

$$\forall \; \eta \in W_h, \; \xi \in V_h.$$

Combining (11.47) and (11.48) we get

$$((T-T_h)u,u)_G = A((S-S_h)u,Su+\xi) + B((S-S_h)u,Tu+\eta)$$

$$+ B(\xi-Su,(T-T_h)u),$$

$$\forall \; \eta \in W_h, \; \xi \in V_h.$$

which, letting $\eta = -T_hu$ and $\xi = S_hu$, yields

$$((T-T_h)u,u)_G = A((S-S_h)u,(S+S_h)u)$$

$$(11.49)$$

$$= 2A((S-S_h)u,Su)-A((S-S_h)u,(S-S_h)u).$$

Now, again using (11.6) and (11.7) we get

$$a((S-S_h)u,Su) = -B((S-S_h)u,Tu)$$

$$= -B((S-S_h)u,Tu - \sum_h(Tu)),$$

197

where $\sum_h Tu$ is the W_h-interpolant of Tu, and hence, using

$$\int_0^1 S_h u [Tu - \sum_h Tu]' dx = 0$$

and (11.6) we have

(11.50) $\qquad A((S-S_h)u, Su) = -B(Su, Tu - \sum_h Tu)$

$$= \int_0^1 bu[Tu - \sum_h Tu]dx.$$

Finally, combining (11.49) and (11.50) we get

(11.51) $\qquad ((T-T_h)u, u)_G = 2 \int_0^1 bu[Tu - \sum_h Tu]dx$

$$- \int_0^1 \frac{|(S-S_h)u|^2}{a} dx$$

$$= 2\lambda^{-1} \int_0^1 bu(u - \sum_h u)dx$$

$$- \int_0^1 \frac{|(S-S_h)u|^2}{a} dx.$$

Now, using (11.50) and (11.46) we get

$$|\lambda - \lambda_h| \leq C\{|\int_0^1 bu(u - \sum_h u)dx| + a_0^{-1} \|(S-S_h)u\|_{L_2}$$

(11.52)
$$+ \|(T-T_h)u\|_{L_2}^2 \}.$$

It remains to estimate the three terms on the right side of (11.52).

Recall that $\sum_h u$ is the W_h-interpolant of u. By a result of Prosdorf and Schmidt [1981] we know that

$$(11.53) \qquad \|u - \sum_h u\|_{L^1} \leq Ch^2 V_0^1(u'),$$

where $V_0^1(u')$ denotes the variation of u'. Recall that u is an eigenfunction of (11.43) with $\|u\|_G = 1$. Since $a(x)$ is of bounded variation, u' will be of bounded variation; in fact

$$(11.54) \qquad V_0^1(u') \leq C,$$

where $C = C(a_0, a_1, b_0, b_1, V_0^1(a), \lambda)$ depends on $a_0, a_1, b_0, b_1, V_0^1(a)$, and λ. Also

$$(11.55) \qquad \|u\|_{L^\infty} \leq C.$$

Using Hölders inequality, together with (11.53) - (11.55), we get

$$(11.56) \qquad |\int_0^1 bu(u - \sum_h u)dx| \leq \|bu\|_{L^\infty} \|u - \sum_h u\|_{L^1}$$
$$\leq Ch^2 V(a),$$

where $C = C(a_0, a_1, b_0, b_1, V_0^1(a), \lambda)$.

Next we consider $\|(S-S_h)u\|_{L_2}$ and $\|(T-T_h)u\|_{L_2}$. It is easily seen that the results in Falk and Osborn [1980] imply

$$(11.57) \qquad \|(S-S_h)u\|_{L_2}, \ \|(T-T_h)u\|_{L_2} \leq C(a_0, a_1, b_0, b_1, \lambda)h.$$

Note that (11.57) shows that $\|T-T_h\|_{GG} \to 0$.

Finally, combining (11.52), (11.56), and (11.57) we have

Theorem 11.5. Suppose λ is an eigenvalue of (11.43) (or of (11.44)) and let λ_h be an the approximate eigenvalue defined by 11.4 with V_h and W_h defined by (11.45). Then

$$(11.58) \qquad |\lambda - \lambda_h| \leq C(a_0, a_1, b_0, b_1, V_0^1(a), \lambda)h^2.$$

The striking feature of estimate (11.58) is that the constant C depends on the bounds a_0, a_1, b_0, and b_1 and on $V_0^1(a)$, but is otherwise independent of $a(x)$ and $b(x)$. This shows that the approximation method is effective for problems with rough coefficients (cf. discussion of alternate variational formulations at the end of Section 3). In fact, the rate of convergence indicated by (11.58) is the same as that for the usual Ritz method for problems with smooth coefficients. (11.58) was proved by Banerjee [1987]. The use of mixed methods for eigenvalue approximation for problems with rough coefficients was first suggested by Nemat-Nasser [1972a, 1972b, 1974]. Rate of convergence estimates for several such mixed method were derived by Babuska and Osborn [1978].

Remark 11.6. It is of interest to note that the variable σ_h can be eliminated from 11.4 in the present context (i.e., with the choices for V, W, H, G, A, B, V_h, and W_h we have made in this subsection) leading to the problem

$$(11.59) \quad \begin{cases} u_h \in W_h \\ \displaystyle\sum_{i=1}^{M} \int_{T_i} a_\tau u_h' \bar{v}' \, dx = \lambda_h \int_0^1 b u_h \bar{v} \, dx, \ \forall \ v \in W_h, \end{cases}$$

where a_τ is a step function with $a_\tau|_{T_i} = \left[\dfrac{\int_{T_i} \frac{dx}{a}}{\text{diam } T_i} \right]^{-1}$, $i = 1, \ldots, M(\tau)$. Thus (11.59) differs from the usual Ritz method only in that the coefficient enters the calculation through its harmonic averages over the subintervals of the mesh instead of through its averages.

Section 12.   Methods Based on One Parameter Families of Varia-
tional Formulations

In our treatment of the membrane problem in Subsection 10.B.,
the trial and test functions satisfied the essential boundary con-
dition $u = 0$ (cf. (10.42)). In fact, if one bases the approxima-
tion method on the usual variational formulation (10.39), one must
impose the boundary condition on the trial and test functions.
To avoid this, methods have been developed that use test and trial
functions that are not required to satisfy essential boundary con-
ditions. (See the discussion of essential and natural boundary
conditions in Section 3.) In this section we discuss two such
methods. They are both based on one parameter family of variation-
al formulations. We will be rather brief and will not explicitly
discuss each of the steps 1), 1'), 2), and 3) of finite element
approximation outlined in Section 10.

A. The Least Squares Method

Consider, as in Subsections 10.B. and 12.B., the membrane
problem

$$
(12.1) \qquad
\begin{cases}
-\Delta u = \lambda u \quad \text{in} \quad \Omega \\
u = 0 \quad \text{on} \quad \Gamma = \partial \Omega,
\end{cases}
$$

where Ω is a bounded, open set with boundary Γ , which, for the
sake of simplicity, we assume to be of class C^∞ . Note that we are
not assuming Ω to be a polygon. (12.1) has eigenvalues

$$
0 < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow \infty
$$

and eigenfunctions

$$
u_1, u_2, \ldots .
$$

We begin by introducing the least squares method for the cor-responding source problem,

$$(12.2) \quad \begin{cases} -\Delta w = f & \text{in } \Omega \\ w = 0 & \text{on } \Gamma, \end{cases}$$

which is usually given the variational formulation (see Remark 12.1 for the reason for using complex functions here),

$$(12.3) \quad \begin{cases} W \in H_0^1(\Omega) \\ \displaystyle\int_\Omega \nabla w \cdot \overline{\nabla v} \ dxdy = \int_\Omega fv \ dxdy, \ \forall \ v \in H_0^1(\Omega). \end{cases}$$

We now give (12.2) a different variational formulation. w solves (12.2) if and only if

$$(12.4) \quad \begin{cases} w \in H^2(\Omega) \\ \displaystyle\int \Delta w \overline{\Delta v} \ dxdy + \rho \int_\Gamma w\overline{v} \ dxdy = -\int_\Omega f\overline{\Delta v} \ dxdy, \\ \qquad\qquad\qquad\qquad \forall \ v \in H^2(\Omega), \ \forall \ 0 < h \le 1, \end{cases}$$

where $\rho = \rho_h \ge 1$ is a parameter that approaches ∞ as $h \to 0$. To pass from (12.2) to (12.4) is immediate. To go from (12.4) to (12.2) we proceed as follows. First take $v \in H^2(\Omega)$ to satisfy

$$\begin{cases} \Delta v = \Delta w + f & \text{in } \Omega \\ v = 0 & \text{on } \Gamma. \end{cases}$$

This chioce for v in (12.4) yields $-\Delta w = f$ in Ω. The equation in (12.4) then becomes

$$\rho \int_\Omega w\overline{v} \ dx = 0, \ \forall \ v \in H^2(\Omega),$$

which implies $w = 0$ on Γ. In (12.4) the boundary conditions

202

$w = 0$ is not explicitly imposed. This is the major advantage of the formulation (12.4) over (12.3) for our purposes. We note that w can also be characterized by an extremal property: the solution w of (12.4) is the unique minimizer of the functional

$$\int_\Omega |-\Delta v - f|^2 \, dxdy + \rho \int_\Gamma |v|^2 ds$$

over $v \in H^2(\Omega)$.

In order to discretize (12.4) we suppose we have a family $\gamma = \{\tau\} = \{\tau_h\}$ of triangulations of $\overline{\Omega'}$, where Ω' is some fixed rectangle containing $\overline{\Omega}$. Then let

$$S_h = S^{p,2}(\tau_h) = \{u \in H^2(\Omega') : u|_T = \text{a polynomial of} \\ \text{degree } p, \ \forall \ T \in \tau_h\}$$

and let S_h consist of the restrictions of functions in $S^{p,2}(\tau_h)$ to Ω. The family S_h satisfies the following approximation result: If $p \geq 5$, then

$$(12.5) \quad \inf_{\chi \in S_h} \sum_{j=0}^{2} h^j \|v - \chi\|_{j,\Omega} \leq Ch^t \|v\|_{t,\Omega}, \quad \text{for} \quad 2 \leq t \leq p+1.$$

See Ciarlet [1978] for a proof of (12.5). Then we define an approximate solution w_h to w by letting w_h be the unique solution to

$$(12.6) \quad \begin{cases} w_h \in S_h \\ \int_\Omega \Delta w_h \overline{\Delta v} \, dxdy + \rho \int_\Gamma w_h \overline{v} \, ds = - \int_\Omega f \, \overline{\Delta v} \, dxdy, \ \forall \ v \in S_h. \end{cases}$$

w_h is called the least squares approximation to w since it can be alternately characterized as the unique minimizer of

$$\int_{\Omega} |-\Delta v - f|^2 dxdy + \rho \int |v|^2 ds$$

over $v \in S_h$. Bramble and Schatz [1970] proposed and analyzed this method for $\rho = \rho_h = h^{-3}$. The also showed $\rho = h^{-3}$ to be the optimal choice for ρ.

Now we return to the eigenvalue problem (12.1). Proceeding in a similar way we see that (12.1) has the variational formulation

$$(12.7) \quad \begin{cases} u \in H^2(\Omega) \\ \int_{\Omega} \Delta u \, \overline{\Delta v} \, dxdy + \rho \int_{\Gamma} u\bar{v} \, ds = -\lambda \int_{\Omega} u\overline{\Delta v} \, dxdy, \; \forall \; v \in H^2(\Omega). \end{cases}$$

(12.7) is then discretized by

$$(12.8) \quad \begin{cases} \lambda_h \text{ complex, } 0 \neq u_h \in S_h \\ \int_{\Omega} \Delta u_h \overline{\Delta v} \, dxdy + \rho \int_{\Gamma} u\bar{v} \, ds = -\lambda_h \int_{\Omega} u_h \overline{\Delta v} \, dxdy, \; \forall \; v \in S_h. \end{cases}$$

(12.8) has eigenpairs $(\lambda_{j,h}, u_{j,h})$, $j = 1, \ldots, N$, where $N = \dim S_h$.

If for $f \in H^0(\Omega)$ we define $Tf = w$ and $T_h f = w_h$, where w and w_h are defined by (12.2) (or (12.4)) and (12.6), respectively, then we easily see that (λ, u) is an eigenpair of (12.1) if and only if $(\mu = \lambda^{-1}, u)$ is an eigenpair of T and (λ_h, u_h) is an eigenpair of (12.8) if and only if $(\mu_h = \lambda_h^{-1}, u_h)$ is an eigenpair of T_h. We will estimate the error in (μ_h, u_h), and thus in (λ_h, u_h), by applying the results in Section 7. T and T_h are clearly compact on $H^0(\Omega)$. We will show $\|T-T_h\| \longrightarrow 0$ in the next paragraph.

In order to apply Theorem 7.3 on $H^0(\Omega)$ we need estimates for

$((T-T_h)u,u)$, $\|(T-T_h)u\|_{0,\Omega}$, and $\|(T-T_h^*)u\|_{0,\Omega}$, where u is an eigenfunction of (12.1) corresponding to the eigenvalue λ (or μ) we are approximating. These estimates are all contained in Bramble and Schatz [1970] (and also in Baker [1973]) for the choice $\rho = h^{-3}$. In their Corollary 4.1 take $\gamma = 3/2$, $\lambda = t-2$, $g = 0$, $\ell = -s$, and $r = p+1$ to get

$$|((T-T_h)\phi,\psi)_{0,\Omega}| \le Ch^{s+t}\|\phi\|_{t-2,\Omega}\|\psi\|_{s,\Omega}, \text{ for}$$

(12.9)
$$0 \le s \le p-3, \ 2 \le t \le p+1.$$

Taking $s = 0$ and $t = 2$ in (12.9) shows that $\|T-T_h\| \to 0$. Now take $s = p-3$ and $t = p+1$ to obtain

(12.10) $\quad |((T-T_h)\phi,\phi)_{0,\Omega}| \le Ch^{2p-2}\|\phi\|_{p-1,\Omega}\|\phi\|_{p-3,\Omega}$,

take $s = 0$ and $t = p+1$ to obtain

$$|((T-T_h)\phi,\psi)_{0,\Omega}| \le Ch^{p+1}\|\phi\|_{p-1,\Omega}\|\psi\|_{0,\Omega},$$

and hence

(12.11) $\quad\quad \|((T-T_h)\phi)\|_{0,\Omega} \le Ch^{p+1}\|\phi\|_{p-1,\Omega}$,

and take $s = p-3$ and $t = 2$ to obtain

$$|((T-T_h)\phi,\psi)_{0,\Omega}| = |(\phi,(T-T_h^*)\psi)_{0,\Omega}| \le Ch^{p-1}\|\phi\|_{0,\Omega}\|\psi\|_{p-3,\Omega},$$

and hence

(12.12) $\quad\quad \|(T-T_h^*)\psi)\|_{0,\Omega} \le Ch^{p-1}\|\psi\|_{p-3,\Omega}$.

Theorem 12.1. Suppose the approximate eigenpairs $(\lambda_{j,h}, u_{j,h})$ are defined by (12.8) with $\rho = h^{-3}$ and suppose the eigenfunctions of (12.1) belong to $H^{p-1}(\Omega)$. Then

(12.13) $\quad\quad\quad |\lambda_{k,h} - \lambda_k| \le Ch^{2p-2}$

and

(12.14) $$\|u_{k,h} - u_k\|_{0,\Omega} \leq Ch^{p+1}.$$

Proof. Let λ_k be any eigenvalue of (12.1) and suppose its geo-metric multiplicity is q, i.e., the geometric multiplicity of $\mu_k = \lambda_k^{-1}$ is q. Since T is selfadjoint, the ascent is one and the algebraic multiplicity of μ_k is also q. q of the $\lambda_{j,h}$ will converge to λ_k. Let $\lambda_{k,h}$ be one of them. Theorem 7.3 can now be applied and (12.13) follows directly from (7.15) and (12.10) - (12.12) since all of the eigenfunctions of (12.1) corresponding to λ_k belong to $H^{p-1}(\Omega)$. (12.14) follows from Theorems 7.1 and 7.4 and (12.11).   ⊔

Remark 12.1. Even though (12.1) is selfadjoint, (12.8) is a non-selfadjoint (finite dimensinal) problem. Thus one needs the gene-ral (not necessarily selfadjoint) theory in Section 7 to analyze the least squares method. The nonselfadjointness of (12.8) is the reason we have used complex function spaces in this analysis.

B. The Penalty Method

We will once more consider the membrane eigenvalue problem (10.38) and assume the boundary Γ of Ω is of class C^∞ (cf. also (12.1)). In Section 10 we gave this problem the variational formulation

(12.15) $$\begin{cases} u \in H_0^1(\Omega) \\ a(u,v) = \lambda b(u,v), \ \forall \ v \in H_0^1(\Omega), \end{cases}$$

where

(12.16a) $$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \ dxdy$$

and

(12.16b)
$$b(u,v) = \int_\Omega uv \, dxdy.$$

Let us replace the boundary condition $u = 0$ on Γ in (10.38) by $u + \psi^{-1}\frac{\partial u}{\partial n} = 0$, i.e., let us consider the problem

(12.17)
$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega \\ u + \psi^{-1}\frac{\partial u}{\partial n} = 0 & \text{on } \Gamma, \end{cases}$$

where $\psi = \psi_h \geq 1$ is a parameter that approaches $+\infty$ as $h \to 0$. It is easily seen that (12.17) has the variational form

(12.18)
$$\begin{cases} u \in H^1(\Omega) \\ a_\psi(u,v) = \lambda b(u,v), \ \forall \ v \in H^1(\Omega), \end{cases}$$

where

(12.19)
$$a_\psi(u,v) = \int_\Omega \nabla u \cdot \nabla u \, dxdy + \psi \int_\Omega uv \, ds.$$

Note that in (12.18), in contrast to (12.15), we have not imposed any constraint on either u or v. This is the case since $u + \psi^{-1}\frac{\partial u}{\partial n} = 0$ is a natural boundary condition (cf. Section 3).

We now estimate the error between the eigenvalues and eigenvectors of (12.15) and (12.18). Toward this end consider the corresponding source problems:

(12.20)
$$\begin{cases} -\Delta u_0 = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma \end{cases}$$

and

(12.21)
$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u + \psi^{-1}\frac{\partial u}{\partial n} = 0 & \text{on } \Gamma. \end{cases}$$

We view (12.21) as an approximation to (12.20). Denoting by u_0

(respectively u_ψ) the solution of (12.20) (respectively (12.21)), we are interested in estimating $u_\psi - u_0$. It is shown in Babuska and Aziz [1973, Section 7.2] that

$$(12.22) \qquad u_\psi = u_0 - \psi^{-1}\xi + \zeta,$$

where ξ is the solution of the problem

$$(12.23) \qquad \begin{cases} -\Delta\xi + \xi = 0 & \text{in } \Omega \\ \xi = \dfrac{\partial u_0}{\partial n} & \text{in } \Omega \end{cases}$$

and ζ is the solution to

$$(12.24) \qquad \begin{cases} \zeta \in H^1(\Omega) \\ a_\psi(\zeta,v) = \psi^{-1}a(\xi,v), \quad \forall\, v \in H^1(\Omega). \end{cases}$$

From (12.16), (12.19), and (12.24) we have

$$\|\zeta\|^2_{H^1(\Omega)} \leq a_\psi(\zeta\ \zeta) = \psi^{-1}(a(\xi,\zeta)$$
$$= \psi^{-1}\|\xi\|_{H^1(\Omega)}\|\zeta\|_{H^1(\Omega)}$$

and hence

$$(12.25) \qquad \|\zeta\|_{H^1(\Omega)} \leq \psi^{-1}\|\xi\|_{H^1(\Omega)}.$$

From (12.23) and (12.25) we obtain

$$(12.26) \qquad \|u_\psi - u_0\|_{H^1(\Omega)} \leq 2\psi^{-1}\|\xi\|_{H^1(\Omega)}.$$

From (12.23) and regularity results for elliptic boundary value problems we get

$$(12.27) \qquad \|\xi\|_{H^1(\Omega)} \leq C\|f\|_{H^0(\Omega)}.$$

Combining (12.26) and (12.27) yields

(12.28)
$$\|u_\varphi - u_0\|_{H^1(\Omega)} \le C\varphi^{-1}\|f\|_{H^0(\Omega)} .$$

If we now introduce the operators T and T_φ on $H^0(\Omega)$ by

$$Tf = u_0$$

and

$$T_\varphi f = u_\varphi ,$$

then (12.28) implies that

(12.29)
$$\|(T-T_\varphi)f\|_{H^0(\Omega)} \le C\varphi^{-1}\|f\|_{0,\Omega} .$$

It is immediate that (λ, u) is an eigenpair of (12.15) if an only if $(\mu = \lambda^{-1}, u)$ is an eigenpair of T; likewise $(\lambda_\varphi, u_\varphi)$ is an eigenpair of (12.18) if and only if $\mu_\varphi = (\lambda_\varphi^{-1}, u_\varphi)$ is an eigenpair of T_φ. It thus follows immediately from Theorems 7.1 - 7.4 and (12.29) that

(12.30)
$$|\lambda_j - \lambda_{\varphi,j}| \le C\varphi^{-1}$$

and

(12.31)
$$\|u_j - u_{\varphi,j}\|_{H^1(\Omega)} \le C\varphi^{-1} ,$$

where (λ_j, u_j) and $(\lambda_{\varphi,j}, u_{\varphi,j})$ denote the eigenpairs of (12.15) and (12.18), respectively. Note that (12.30) and (12.31) are estimates of the same order for both the eigenvalue and eigenvector errors. This is in contrast to approximations we have analyzed previously in this article. An analysis of a one-dimensional model problem shows that, for the type of approximations we are considering, the eigenvalue and eigenvector error is, indeed, of the same order.

Next we consider the problem (12.18) and approximate it by a

209

finite element method, letting the resulting eigenpairs be $(\lambda_{\psi,j,h}, u_{\psi,j,h})$. Since u and v in (12.18) are taken in $H^1(\Omega)$, we need not impose any boundary condition on the trial and test space S_h. If one now analyzes the error in the finite element approximation of (12.18), selects ψ so that the error in passing from (12.15) to (12.18) is of the same magnitude of that incurred in the finite element approximation of (12.18), and then combines the error estimates (12.30) and (12.31) with those for the finit element approximations of (12.18), one obtains estimates for the difference between (λ_j, u_j) and $(\lambda_{\psi,j,h}, u_{\psi,j,h})$. We stress that the $(\lambda_{\psi,j,h}, u_{\psi,j,h})$ are calculated from a matrix eigenvalue problem corresponding to trial and test spaces that are not required to satisfy the essential boundary condition for the membrane problem (12.1). The approximation method we have outlined is referred to as the penalty method.

We refer the reader to Babuska and Aziz [1973, Section 7.2] for a detailed analysis of the penalty method for the source problem. Estimates for the errors in eigenvalue approximation can be easily derived from the corresponding source problem estimates by means of Theorems 7.1 - 7.4. Because this application of these error estimates to the eigenvalue problem is similar to those discussed above and raises no new issues, we will not give a formal statement of the results.

Remark 12.2. If Ω is a polygon, then the choice $\psi = \infty$ corresponds to satisfying the boundary condition on $\partial\Omega$, i.e., constraining $S^p(\tau)$ to be $S_0^p(\tau)$, and the resulting method is identical with that discussed in the Subsection 10.B. If Γ is not

polygonal, then $v = \infty$ will lead to the constraint $S^p(\tau) = \tilde{S}_0^p(\tau)$, where $\tilde{S}_0^p(\tau)$ consists of those $u \in S^p(\tau)$ which are zero on every triangle which intersects Γ. The finite element solution then solves the problem on $\tilde{\Omega}$ instead of Ω, where $\tilde{\Omega}$ consists of the union of all triangles which do not intersect Γ. Sometimes the mesh is constructed so that $\Omega - \tilde{\Omega}$ is as small as possible by interpolating Γ by straight lines.

Remark 12.3. In practical computation (codes) the penalty method (or some equivalent method) is often also used when Ω is a polygon by taking ψ to be very large (say $\psi = 10^8$). This is just a way of imposing the essential boundary conditions in the code.

Remark 12.4. The least squares method and penalty method are seldom used as a way to treat essential boundary conditions on a curved boundary because of the difficulty in the computation of $a_\psi(u,v)$, which requires area integrations over triangles which intersect the boundary. The usual approach is to use curvilinear elements, which allow exact satisfaction of the boundary condition in a similar way as when the domain is polygonal (cf. Remark 12.3).

Let us end this section by noting some similarities and differences in the least squares and penalty methods.

- Both methods circumvent essential boundary conditions by reformulating the original problem in terms of a one parameter family of variational formulations. In both methods, the optimal value of the parameter depends on the mesh, i.e., on h.

- With the least squares method, the optimal value of the parameter $(p = h^{-3})$ is independent of the solution.

211

This is related to the fact that the alternate variational formulation characterizes the solution exactly for any value of the parameter. In the case of the penalty method, the optimal value of the parameter depends on the mesh and the smoothness of the solution or the eigenfunction. This is related to the fact that the exact solution does not exactly satisfy the one parameter family of formulations for any value of the parameter $\psi \neq +\infty$.

- The least squares method employs C^1-elements (i.e., subspaces of $H^2(\Omega)$), whereas the penalty method employs C^0-elements (i.e., subspaces of $H^1(\Omega)$). As we have previously noted, C^0-elements are easier to construct than C^1-elements.

Section 13.   Concluding Remarks

A. We have illustrated the application of the general theory that
was presented in Chapter II by considering several important model
problems. It should be clear from the analysis of these model
problems how to treat a wide variety of problems. We have seen,
however, that the application of the general theory to a concrete
problem may require subtle analysis.

B. In Sections 10, 11, and 12 we have illustrated the main
approach to finite element approximation of eigenvalue problems.
We have seen that there are many available methods and that their
basic theoretical properties can be established as an application
of the results in Chapter II. Nevertheless, the implementation of
these methods raises many other important questions; although we
cannot address these questions in detail, we now mention some of
them.

1) Which method is most effective for a specific problem?
What is the goal of the computation? We remark that sometimes high
accuracy is achieved for eigenvalue approximation, but that only
low accuracy is obtained for the approximation of other important
quantities such as the stresses, moments, or shear forces.

2) What types of meshes or adaptive mesh procedures are
desirable? How should the quality of the computed results be
assessed a posteriori? For a survey of results in this direction,
see Noor and Babuska [1987].

3) Which matrix eigenvalue solvers should be used? What
computer architecture is desirable (sequential, parallel)?

These questions are, of course, not restricted to eigenvalue

computation. They also arise with finite element computation of source problems. Some of these questions may be addressed in other articles in this Handbook.

C. The Ritz method, which was discussed in Section 10, is most easily analyzed with the results of Section 8, specifically with (8.44) - (8.46). Note that because of (8.32), (8.11) is satisfied with $\beta = \alpha$ and thus the major requirement on S_h is that it have good approximation properties.

D. We have seen in Remark 11.2 that mixed methods for eigenvalue approximation have the form of (8.10) and (8.14). Thus, if a method satisfies the hypotheses of Section 8, specifically (8.1), (8.2), (8.3), (8.6), (8.11), and (8.13), then the method can also be analyzed with the results of Section 8. Most mixed methods, however, fail to satisfy at least one of these hypotheses, and we thus cannot rely on the results of Section 8. We now comment on two of the examples discussed in Section 11 in regard to which results in Chapter II their analysis is based on.

1) Consider first the mixed method discussed in Subsection 11.A. for the membrane problem. It is easily seen that the variational formulation (11.24) satisfies (8.1), (8.2), and (8.3), but that it does not satisfy (8.6). In Section 8, assumption (8.6) is used to show that the operator T defined by (8.8) is compact. For our example, for $(f,g) \in H(\text{div},\Omega) \times H^0(\Omega)$,

$$T(f,g) = (\sigma, u),$$

where u solves

$$\begin{cases} u \in H_0^1(\Omega) \\ -\Delta u = g \quad \text{in} \quad \Omega \end{cases}$$

214

and $\sigma = \nabla u$, and, by noting in particular the dependence of σ on g, we see that $T : H(\text{div},\Omega) \times H^0(\Omega) \longrightarrow H(\text{div},\Omega) \times H^0(\Omega)$ is not compact. Since T is not compact, T_h, as defined by (8.16), cannot converge to T in norm. Because of these facts, the results of Section 7 do not apply (to this T). The analysis that we used for this problem (cf. Theorem 11.1) is based on Section 7 and circumvents this difficulty by using a different operator, namely $T : H^0(\Omega) \longrightarrow H^0(\Omega)$ defined by $Tg = u$ (cf. 11.8c).

As mentioned in Remark 7.7, results for noncompact operators which parallel those in Section 7 have been proved by Descloux, Nassif, and Rappaz [1978a, 1978b], and one can, if fact, use them to derive the estimates we obtained in Subsection 11.B, specifically (11.32) and (11.33). We will not present the details of this analysis but will comment briefly on the applicability of the results of Descloux, Nassif, and Rappaz [1978a, 1978b] to our problem.

For their results, T is not required to be compact and T_h is assumed to converge to T in the sense that

(13.1)
$$\inf_{(\tau,\eta) \in S_h = V_h \cdot W_h} \|(\sigma,u) - (\tau,\eta)\|_H \longrightarrow 0 \quad \text{for each} \quad (\sigma,u) \in H(\text{div},\Omega) \cdot H^0(\Omega)$$

and

(13.2)
$$\|T_h - T\|_h = \sup_{(f,g) \in V_h \cdot W_h} \frac{\|(T_h - T)(f,g)\|_{H(\text{div},\Omega) \cdot H^0(\Omega)}}{\|(f,g)\|_{H(\text{div},\Omega) \cdot H^0(\Omega)}} \longrightarrow 0, \quad \text{as } h \longrightarrow 0.$$

With V_h and W_h defined as in (11.31), (13.1) follows from the

215

approximation result in Raviart and Thomas [1977].

We now verify (13.2), which is central point in any application of the results of Descloux, Nassif, and Rappaz [1978a, 1987b]. For $(f,g) \in V_h \times W_h$, let $(\sigma,u) = T(f,g)$ and $(\sigma_h,u_h) = T_h(f,g)$, where T and T_h are defined by (8.8) and (8.16), respectively, for the problem discussed in Subsection 11.B. We know that $u \in H^1(\Omega)$, $-\Delta u = g$, and $\sigma = \nabla u$, and hence div $\sigma = g$. Also, if $g \in W_h$ it is easily seen that div $\sigma_h = -g$. Thus

$$\|(T_h-T)(f,g)\|_{H(\mathrm{div},\Omega) \times H^0(\Omega)} = \|(\sigma_h,u_h)-(\sigma,u)\|_{H(\mathrm{div},\Omega) \times H^0(\Omega)}$$

$$= \left[\|\mathrm{div}\, \sigma_h - \mathrm{div}\, \sigma\|^2_{0,\Omega} + \|\sigma_h - \sigma\|^2_{H^0(\Omega)} \right.$$

(13.3)
$$\left. + \|u_h - u\|^2_{0,\Omega} \right]^{1/2}$$

$$= \left[\|\sigma_h - \sigma\|^2_{H^0(\Omega)} + \|u_h - u\|^2_{0,\Omega} \right]^{1/2},$$

$$\text{for } (f,g) \in V_h \times W_h.$$

From the results in Falk and Osborn [1980] we have

(13.4a)
$$\|\sigma_h - \sigma\|_{H^0(\Omega)} \leq Ch\|u\|_{2,\Omega}$$
$$\leq Ch\|g\|_{0,\Omega}$$

and

(13.4b)
$$\|u_h - u\|_{0,\Omega} \leq Ch^2\|u\|_{2,\Omega}$$
$$\leq Ch^2\|g\|_{0,\Omega}.$$

Combining (13.3) and (13.4) we get

$$\| (T_h - T)(f,g) \|_{H(\text{div},\Omega) \times H^0(\Omega)} \leq Ch \| g \|_{0,\Omega}$$

(13.5) $$\leq Ch \| (f,g) \|_{H(\text{div},\Omega) \times H^0(\Omega)},$$

$$\text{for} \quad (f,g) \in V_h \times W_h.$$

(13.2) follows directly from (13.5).

2) Consider next the method discussed in Subsection 11.C for the vibrating plate problem. The variational formulation (11.38) for the problem does not satisfy (8.2) and (8.11). Note that the method was analyzed by means of Theorem 11.1 which is based on Theorem 7.3

Remark 13.1. The fact that many mixed approximation methods fail to satisfy the usual hypotheses (cf. Babuška [1971, 1973] and Brezzi [1974]) for variational approximation methods is an issue for the approximation of source problems as well as eigenvalue problems. The abstract results in Falk and Osborn [1980] have as their main application the analysis of mixed methods which fail to satisfy the usual hypotheses for variational approximation methods. In this connection see also Babuška, Osborn, and Pitkäranta [1980], where problem (11.38) is reformulated in terms of alternate spaces with alternate (mesh dependent) norms so as to satisfy the usual hypotheses.

Selective Bibliography

Adams, R.A. [1975]: Sobolev Spaces, Academic Press, New York.

Agmon, S. [1965]: Lectures on Elliptic Boundary Value Problems, Van Nostrand Mathematical Studies No. 2, Princeton, N.J.

Aronszajn, N. [1948]: The Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues. I; Operators in a Hilbert space. II: Differential operators, Proc. Nat. Acad. Sci. U.S.A. 34, 474-480, 594-601.

Aronszajn, N. [1949-1950]: The Rayleigh-Ritz and the Weinstein methods for approximation of eigenvalues, Tech. Reps, 1-4, Oklahoma A&M College, Stillwater.

Aronszajn, N.; Weinstein, A. [1942]: On the unified theory of eigenvalues of plates and membranes, Amer. J. Math. 64, 623-645.

Babuska, I. [1971]: Error bounds for finite element method, Numer. Math. 16, 322-333.

Babuska, I.; Aziz, A. [1973]: Survey lectures on the mathematical foundations of the finite element method" in The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations (Editor: A.K. Aziz), 5-359, Academic Press, New York.

Babuska, I.; Guo, B.Q. [1987a]: Regularity of the solution of elliptic problems with piecewise analytic Data. Part 1: boundary value problems for linear elliptic equations of the second order, SIAM J. Math. Anal. To appear.

Babuska, I.; Guo, B.Q. [1987b]: The h-p version of finite element method with curved boundary, SIAM J. Math. Anal. To appear.

Babuska, I.; Guo, B.Q. [1987c]: The h-p version of finite element method for the problems with non-homogeneous Dirichlet data, IPST Tech. Note, Univ. of Maryland, College Park.

Babuska, I.; Miller, A. [1984]: The post-processing approach in the finite element method. Part 2. The calculation of stress intensity factors, Internat. J. Numer. Math. Engrg. 20, 1111-1129.

Babuska, I.; Osborn, J. [1978]: Numerical treatment of eigenvalue problems for differential equations with discontinuous coefficients, Math. Comp. 32, 991-1023.

Babuska, I.; Osborn, J. [1983]: Generalized finite element methods and their relation to mixed methods, SIAM J. Numer. Anal. 20, 510-536.

Babuska, I.; Osborn, J. [1985]:   Finite element methods for the
     solution of problems with rough data, in Singularities and
     Constructive Methods for Their Treatment (Editors: P.
     Grisvard, W. Wendland, and J.R. Whitman), Springer-Verlag
     Lecture Notes in Mathematics 1121, Springer-Verlag, Berlin-
     New York, 1-18.

Babuska, I.; Osborn, J. [1987]:   Estimates for the errors in eigen-
     value and eigenvector approximation by Galerkin methods, with
     particular attention to the case of multiple eigenvalues,
     SIAM J. Numer. Anal.   To appear.

Babuska, I.; Osborn, J.; Pitkaranta, J [1980]:   Analysis of mixed
     methods using mesh dependent norms, Math. Comp. 35, 1039-1062.

Babuska, I.; Scapolla, T. [1987]:   The computational aspects of
     the h,p and h-p versions of finite element method,
     Proceedings of the 6th IMACS Symposium.

Babuska, I.; Suri, M. [1987a]:   The optimal convergence rate of
     the p-version of finite element method, SIAM J. Numer. Anal.
     To appear.

Babuska, I.; Suri, M. [1987b]:   The h-p version of the finite
     element method with quasiuniform meshes, R.A.I.R.O., Mathema-
     tical Modelling and Numerical Analysis.   To appear.

Babuska, I.; Rank, E. [1982]:   An expert-system-like feedback
     approach in the h-p version of finite element method,
     Finite Elements in Analysis and Design.

Baker, G. [1973]:   Simplified proofs of error estimates for the
     least squares method for Dirichlet's problem, Math. Comp. 27,
     229-235.

Banerjee, U. [1987]:   Approximation of eigenvalues of differential
     equations with non-smooth coefficients, R.A.I.R.O., Mathemati-
     cal Modelling and Numerical Analysis.   To appear.

Bazley, N.W. [1959]:   Lower bounds for eigenvalues with applica-
     tions to the helium atom, Proc. Nat. Acad. Sci. U.S.A. 45,
     850-853.

Bazley, N.W. [1961]:   Lower bounds for eigenvalues, J. Math. Mech.
     10, 289-308.

Bazley, N.W.; Fox, D.W. [1961a]:   Truncation in the method of
     intermediate problems for lower bounds to eigenvalues, J.
     Res. Nat. Bur. Standards 65B, 105-111.

Bazley, N.W.; Fox, D.W. [1961b]:   Lower bounds for eigenvalues of
     Schrodinger's equation, Phys. Rev. 124, 483-492.

Bazley, N.W.; Fox, D.W. [1963]: Lower bounds for energy levels of
    molecular systems, J. Mathematical Phys. 4, 1147-1153.

Birkhoff, G.; De Boor, C.; Swartz, B; Wendroff, B. [1966]:
    Rayleigh-Ritz approximation by piecewise cubic polynomials,
    SIAM J. Numer. Anal. 3, 188-203.

Bramble, J.H.; Osborn, J.E. [1973]: Rate of convergence estimates
    for nonselfadjoint eigenvalue approximations, Math. Comp. 27,
    525-549.

Bramble, J.H.; Schatz, A.H. [1970]: Rayleigh-Ritz-Galerkin methods
    for Dirichlet's problem using subspaces without boundary con-
    ditions, Comm. Pure Appl. Math. 23, 653-675.

Brezzi, F.; Raviart, P. [1978]: Mixed finite element methods for
    4th order elliptic equations, Topics in Numerical Analysis
    III (Editor: J. Miller), Academic Press, New York.

Brezzi, F. [1974]: On the existence uniqueness and approximation
    of saddle-point problems arising from Lagrangian multipliers,
    R.A.I.R.O., R2 8, 129-151.

Butzer, P.L.; Berens, H. [1967]: Semi-Groups of Operators and
    Approximation, Lecture Notes in Mathematics 145, Springer-
    Verlag, Berlin-New York.

Canuto, C. [1978]: Eigenvalue approximations by mixed methods,
    R.A.I.R.O. Anal. Numer. 12, 27-50.

Chatelin, F. [1973]: Convergence of approximate methods to compute
    eigenelements of linear operators, SIAM J. Numer. Anal. 10,
    939-948.

Chatelin, F. [1981]: The spectral approximation of linear opera-
    tors with applications to the computation of eigenelements of
    differential and integral operators, SIAM Rev. 23, 495-522.

Chatelin, F. [1983]: Spectral Approximations of Linear Operators,
    Academic Press, New York.

Chatelin, F.; Lemordant, J. [1978]: Error bounds in the approxi-
    mation of eigenvalues of differential and integral operators,
    J. Math. Anal. Appl. 62, 257-271.

Ciarlet, P. [1978]: The Finite Element Method for Elliptic Prob-
    lems, North Holland, Amsterdam.

Ciarlet, P.; Raviart, P. [1974]: A mixed finite element method for
    the biharmonic equation, Symposium on Mathematical Aspects of
    Finite Elements in Partial Differential Equations (Editor: C.
    De Boor), 125-143, Academic Press, New York.

Collatz, L. [1948]: Eigenwertprobleme und Ihre Numerisch
    Behandlung, Chelsea, New York.

Courant, R. [1927]: Über direkte Methoden in der Variationsrech-
nung und über verwandte Fragen, Math. Ann. 97, 711-736.

Courant, R. [1943]: Variational methods for the solution of prob-
lems of equilibrium and vibrations, Bull. Amer. Math. Soc. 49,
1-23.

Courant, R.; Hilbert, D. [1953]: Methods of Mathematical Physics,
Vol. 1. Wiley (Interscience), New York.

Cullum, J.; Willoughby, R.A. [1985]: Lanczos Algorithms for Large
Symmetric Eigenvalue Computations, Vol. I: Theory, Vol. II:
Programs, Birkhäuser, Boston.

Descloux, J.; Nassif, N.; Rappaz, J. [1978a]: On spectral approx-
imation. Part 1. The problem of convergence, R.A.I.R.O.
Anal. Numer. 12, 97-112.

Descloux, J., Nassif, N.; Rappaz, J. [1978b]: Part 2, Error
estimates for the Galerkin method, R.A.I.R.O. Anal. Numer.
12, 113-119.

Dunford, N.; Schwartz, J.T. [1958]: Linear Operators, Part I:
General Theory, Wiley (Interscience), New York.

Dunford, N.; Schwartz, J.T. [1963]: Linear Operators, Part II:
Spectral Theory, Selfadjoint Operators in Hilbert Spaces,
Wiley (Interscience), New York.

Falk, R. [1978]: Approximation of the biharmonic equation by a
mixed finite element method, SIAM J. Numer. Anal. 15, 556-567.

Falk, R. ; Osborn, J. [1980]: Error estimates for mixed methods,
R.A.I.R.O. Anal. Numer. 14, 249-277.

Fix, G.J. [1973]: Eigenvalue approximation by the finite element
method, Advances in Math. 10, 300-316.

Fix, G.J. [1972]: Effects of quadrature errors in finite element
approximation of steady state, eigenvalue, and parabolic
parabolic problems, in The Mathematical Foundations of the
Finite Element Method with Applications to Partial Differen-
tial Equations (Editor: A.K. Aziz) 525-556, Academic Press,
New York.

Glowinski, R. [1973]: Approximations externes par élements finis
de Lagrange d'ordre un et deux, du probléme de Dirichlet pour
l'opérateur bihormonique, Méthodes iteratives de résolutions
des problèms approchés, in Topics in Numerical Analysis
(Editor: J.J.H. Miller) 123-171, Academic Press, New York.

Grisvard, P. [1985]:   Elliptic Problems in Nonsmooth Domain,
    Pitman Publishing, Inc.

Gui, W.; Babuska, I. [1986]:   The h,p,h-p versions of finite
    element methods in one dimension, Numer. Math. 49, 577-683.

Guo, B.Q.; Babuska, I. [1986]:   The h-p version of finite element
    method, Part 1:   The basic approximation results, Comp. Mech.
    1, 21-41, Part 2:   General results and application, Comp.
    Mech. 1, 203-220.

Grigorieff, R.D. [1975a]:   Diskrete approximation von eigenwert-
    problemen, I.   Qualitative konvergenz.   Numer. Math. 24,
    355-374.

Grigorieff, R.D. [1975b]:   Diskrete approximation von eigenwert-
    problemen, II.   Konvergenzordnung.   Numer. Math. 24, 415-433.

Grigorieff, R.D. [1975c]:   Diskrete approximation von eigenwert-
    problemen, III.   Asymptotische entwicklung.   Numer. Math. 25,
    79-97.

Hubbard, B.E. [1961]:   Bounds for eigenvalues of the free and fixed
    membrane by finite difference methods, Pacific J. Math. 11,
    559-590.

Hubbard, B.E. [1962]:   Eigenvalues of the non-homogeneous rec-
    tangular membrane by finite difference methods, Arch. Rational
    Mech. Anal. 9, 121-133.

Herrmann, L. [1967]:   Finite element bending analysis for plates,
    J. Eng. Mech., Div. ASCE EM5, 93, 49-83.

Hersch, J. [1955]:   Equations differentielles et fonctions des
    cellules, C.R. Acad. Sci., Paris 240, 1602-1605.

Hersch, J. [1963]:   Lower bounds for all eigenvalues by cell func-
    tions: a refined form of H.F. Weinberger's method, Arch.
    Rational Mech. Anal. 12, 361-366.

Ishihara, K. [1977]:   Convergence of the finite element method
    applied to the eigenvalue problem $\Delta u + \lambda u = 0$, Publications
    of the Research Institute for Mathematical Sciences, Kyoto
    University 13, 48-60.

Ishihara, K. [1978a]:   The buckling of plates by the mixed finite
    element method, Memoirs of Numerical Mathematics 5, 73-82.

Ishihara, K. [1978b]:   A mixed finite element method for the bihar-
    monic eigenvalue problem of plate bending, Publications of
    the Research Institute for Mathematical Sciences, Kyoto Uni-
    versity 14, 399-414.

Johnson, C. [1973]:   On the convergence of a mixed finite element
    method for plate bending problems, Numer. Math. 21, 43-62.

Kato, T. [1958]:   Perturbation theory for nullity, deficiency and
     other quantities of linear operators, J. Analyse Math. $\underline{6}$,
     261-322.

Kato, T. [1966]:   Perturbation Theory for Linear Operators, Lecture
     Notes in Mathematics 132, Springer-Verlag, Berlin-New York.

Kellogg, R.B.; Osborn, J.E.:   A regularity result for the Stokes
     problem in a convex polygon, J. Functional Analysis. 21,
     397-431.

Knops, R.J.; Payne, L.E. [1971]:   Uniqueness Theorems in Linear
     Elasticity, Springer-Verlag, Berlin.

Kolata, W. [1978]:   Approximation of variationally posed eigenvalue
     problems, Numer. Math. 29, 159-171.

Kondratev, V.A. [1968]:   Boundary Problems for elliptic equations
     with conical or angular points, Trans. Moscow Math. Soc. 16
     (1967), translated by Am. Math. Soc.

Kreiss, H.O. [1972]:   Difference approximation for boundary and
     eigenvalue problems for ordinary differential equations,
     Math. Comp. 26, 605-624.

Kufner, A. [1985]:   Weighted Sobolev Spaces, Wiley, New York.

Kuttler, J.R. [1970a]:   Finite difference approximations for
     eigenvalues of uniformly elliptic operators, SIAM J. Numer.
     Anal. 7, 206-232.

Kuttler, J.R. [1970b]:   Upper and lower bounds for eigenvalues by
     finite differences, Pacific J. Math. 35, 429-440.

Lax, P.D., Milgram, A.N. [1954]:   Parabolic equations, Annals of
     Mathematics Studies No. 33, 167-190, Princeton University
     Press, Princeton.

Mercier, B. [1974]:   Numerical solution of the biharmonic problems
     by mixed finite elements of Class C^0, Boll. Un. Mat. Ital.
     10, 133-149.

Mercier, B.; Osborn, J.; Rappaz, J.; Raviart, P. [1981]:   Eigen-
     value approximation by mixed and hybrid methods, Math. Comp.
     36, 427-453.

Mercier, B.; Rappaz, J. [1978]:   Eigenvalue approximation via non-
     conforming and hybrid finite element methods, Rapport du
     Centre de Mathematiques Appliquees, Ecole Polytechnique,
     Palaiseau, France.

Merigot, M. [1974]:   Regularite des functions properes du Laplacien
     dans un cone, C.R. Acad. Sci. Paris, Ser. A 279, 503-505.

Morrey, C.B. [1966]: Multiple Integrals in Calculus of Variations, Berlin-Heidelberg-New York, Springer-Verlag.

Necas, J. [1967]: Les Méthodes Directes en Théorie des Équations Elliptiques, Paris, Masson.

Necas, J.; Hlavacek, I. [1970]: On inequalities of Korn's type. II. Applications to linear elasticity, Arch. Rational Mech. Anal. 36, 312-334.

Nemat-Nasser, S. [1972a]: General variational methods for elastic waves in composites, J. Elasticity 2, 73-90.

Ner .-Nasser, S. [1972b]: Harmonic waves in layered composites, J. Appl. Mech. 39, 850-852.

Nemat-Nasser, S. [1974]: General Variational Principles in Non-linear and Linear Elasticity with Applications, Mechanics Today 1, pp. 214-261 Pergamon Press, New York.

Noor, A.K. [1985]: Books and Monographs on finite Element Technology, Finite Elements in Analysis and Design 1, 101-111.

Noor, A.K.; Babuska, I. [1987]: Quality Assessment and Control of finite element solutions, Finite Elements in Analysis and Design. To appear.

Oden, J. [1973]: Some Contributions to the Mathematical Theory of Mixed Finite Element Approximations, in Theory and Practice in Finite Element Structural Analysis, pp. 3-23 University of Tokyo Press.

Oden, J.T.; Carey, G.T. [1982]: Finite Elements Mathematical Aspects, Vol. IV, Prentice-Hall, Englewood Cliffs, N.J.

Oden, J.T.; Reddy, J.N. [1976]: An Introduction to the Mathematical Theory of Finite Elements, Wiley, New York.

Osborn, J.E. [1975]: Spectral approximation for compact operators, Math. Comp. 26 712-725.

Osborn, J.E. [1979]: Eigenvalue approximations by mixed methods, in Advances in Computer Methods for Partial Differential Equations - III, Proceedings of the Third IMACS International Symposium on Computer Methods for Partial Differential Equations held at Lehigh University (Editors: R. Vichnevetsky and R. Stepleman), 158-161.

Osborn, J.E. [1981]: The numerical solution of differential equations with rough coefficients, in Advances in Computer Methods for Partial Differential Equations - IV, Proceedings of the Fourth IMACS International Symposium on Computer Methods for Partial Differential Equations held at Lehigh University (Editors: R. Vichnevetsky and R. Stepleman), 9-13.

Parlet, B.N. [1980]: The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, New Jersey.

Polya, G. [1952]: Sur une interprétation de la méthode des differences finie qui peut fournir des bornes supérieures ou inférieures, C.R. Acad. Sci., Paris 235, 995-997.

Prodi, G. [1962]: Theoremi di tipo locale per il sistema de Navier-Stokes e stabilità della soluzioni stationarie, Rend. Sem. Mat. Univ. Padua 32, 374-397.

Prosdorf, S.; Schmidt, G. [1981]: A finite element collocation method for singular integral equations, Math. Nachr. 100 33-60.

Raviart, P.A.; Thomas, J.M., A Mixed Finite Element Method for 2nd Order Elliptic Problems, Lecture Notes in Math., Vol. 606, 292-315, Springer-Verlag, Berlin and New York, 1977.

Scholz, R. [1976]: Approximation von Sattelpunkten mit Finiten Elemented, Tagungsband, Bonn. Math. Schr. 89, 53-66.

Strang, G.; Fix, G.J. [1973]: An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, New Jersey.

Stummel, F. [1979]: Approximation methods for eigenvalue problems in elliptic differential equations, in Numerik und Anwendungen ron Eigenwertaufgaben und Verzweigungsproblemen (Editors: E. Bohl, L. Collatz, and K.P. Hadeler), pp. 133-165, Birkhäuser, Basel.

Synge, J.L.; Griffith, B.A. [1959]: Principles of Mechanics, 34d Ed., McGraw-Hill.

Szabo, B.A. [1985]: PROBE: Theoretical Manual, NOETIC Tech., St. Louis.

Szabo, B.A. [1986]: Mesh Design for the p-version of the Finite Element Method, Comput. Methods Appl. Math. Engr. 55, 181-197.

Szabo, B.A.; Babuska, I. [1986]: Computation of the amplitude of stress, singular terms for cracks and reentrant corners, Rep. WU/CCM-86/1, Center for Computational Mechanics, Washington University, St. Louis.

Vainikko, G.M. [1964]: Asymptotic error bounds for projection methods in the eigenvalue problem, U.S.S.R. Comput. Math. and Math. Phys. 4, 1964, 9-36.

Vainikko, G.M. [1967]: Rapidity of convergence of approximation methods in the eigenvalue problem," U.S.S.R. Comput. Math. and Math. Phys. 7, 18-32.

Vainikko, G.M. [1970]: On the rate of convergence of certain approximation methods of Galerkin type in an eigenvalue problem, Amer. Math Soc. Transl. 36, 249-259.

Weinberger, H.F. [1952a]: Error estimation in the Weinstein method for eigenvalues, Proc. Amer. Math. Soc. 3, 643-646.

Weinberger, H.F. [1952b]: An optimum problem in the Weinstein method for eigenvalues, Pacific J. Math. 2, 413-418.

Weinberger, H.F. [1956]: Upper and lower bounds for eigenvalues by finite difference methods, Comm. Pure. Appl. Math. 9, 613-623 (also, Proc. Conference on Partial Differential Equations, Berkeley, Calif., 1955).

Weinberger, H.F. [1958]: Lower bounds for higher eigenvalues by finite difference methods, Pacific J. Math. 8, 339-368.

Weinberger, H.F. [1959]: A theory of lower bounds for eigenvalues, Tech Note BN-183, Inst. for Fluid Dynamics and Applied Mathematics, Univ. of Maryland, College Park.

Weinberger, H.F. [1960]: Error bounds in the Rayleigh Ritz approximation of eigenvectors, J. Res. Nat. Bur. Standards 64B, 217-225.

Weinberger, H.F. [1974]: Variational Methods for Eigenvalue Approximation, Regional Conference Series in Appl. Math. 15, Society for Industrial and Applied Mathematics, Philadelphia.

Weinstein, A. [1935]: Sur la stabilité des plaques encastrées, C.R. Acad. Sci. Paris 200, 107-109.

Weinstein, A. [1937]: Étude des spectres des équations aux dérivées partielles de la théorie des plaques élastiques, Mémor. Sci. Math., no. 88, Gauthier-Villars, Paris.

Weinstein, A. [1953]: Variational methods for the approximation and exact computation of eigenvalues, Nat. Bur. Standards Appl. Math. Ser. 29, 83-89.

Weinstein, A. [1963]: On the Sturm-Liouville theory and the eigenvalues of intermediate problems, Numer. Math. 5, 238-245.

Weinstein, A.; Stenger, W. [1972]: Methods for Intermediate Problems for Eigenvalues, Academic Press, New York.

Weinstein, D.H. [1934]: Modified Ritz method, Proc. Nat. Acad. Sci. U.S.A. 20, 529-532.

Woznikowski, H.A. [1985]: A survey of information-based complexity, Journal of Complexity 1, 11-44.

Subject Index

Author Index

Aziz, K., 74, 75, 76, 81, 110, 208, 210

Adams, R.A., 27

Agmon, S., 24, 36, 47

Aronszajn, N., 171

Babuška, I., 36, 48, 74, 75, 76, 81, 89, 110, 124, 129, 137, 138,
141, 142, 143, 144, 148, 162, 167, 170, 172, 200, 208, 210,
213, 217

Baker, G., 205

Banerjee, U., 200

Bazley, N.W., 171

Berens, H., 27

Birkhoff, G., 110

Bramble, J.H., 63, 64, 204, 205

Brezzi, F., 172, 217

Butzer, P.L., 27

Canuto, C., 194

Carey, G.T., 171

Chatelin, F., 63

Ciarlet, P., 123, 124, 170, 172, 191

Collatz, L., 171

Courant, R., 2, 57, 86, 170

Cullum, J., 143

De Boor, C., 110

Descloux, J., 67, 73, 215, 216

Dunford, N., 43, 59

Stenger, W., 86, 171

Strang, G., 171

Stummel, F., 63

Suri, M., 124, 137, 148

Swartz, B., 110

Synge, J.L., 57

Szabo, B.A., 138, 143, 170

Thomas, J.M., 172, 186, 187, 216

Vainikko, G.M., 63

Weinberger, H.F., 86, 170, 171

Weinstein, A., 86, 171

Weinstein, D.A., 171

Wendroff, B., 110

Willoughby, R.A., 143

Woznikowski, H.A., 167

The <u>Laboratory for Numerical analysis</u> is an integral part of the Institute for Physical Science and Technology of the University of Maryland, under the general administration of the Director, Institute for Physical Science and Technology. It has the following goals:

o   To conduct research in the mathematical theory and computational implementation of numerical analysis and related topics, with emphasis on the numerical treatment of linear and nonlinear differential equations and problems in linear and nonlinear algebra.

o   To help bridge gaps between computational directions in engineering, physics, etc., and those in the mathematical community.

o   To provide a limited consulting service in all areas of numerical mathematics to the University as a whole, and also to government agencies and industries in the State of Maryland and the Washington Metropolitan area.

o   To assist with the education of numerical analysts, especially at the postdoctoral level, in conjunction with the Interdisciplinary Applied Mathematics Program and the programs of the Mathematics and Computer Science Departments. This includes active collaboration with government agencies such as the National Bureau of Standards.

o   To be an international center of study and research for foreign students in numerical mathematics who are supported by foreign governments or exchange agencies (Fulbright, etc.)

<u>Further information</u> may be obtained from Professor I. Babuška, Chairman, Laboratory for Numerical Analysis, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

END

FILMED

FEB. 1988

DTIC